European Journal of Political Economy 79 (2023) 102439

Contents lists available at ScienceDirect

European Journal of Political Economy

journal homepage: www.elsevier.com/locate/ejpe

Reputation and perverse transparency under two concerns ☆

Ascensión Andina-Díaz [a,*], José A. García-Martínez [b]

[a] Dpto. Teoría e Historia Económica, Universidad de Málaga, Spain
[b] Dpto. Estudios Económicos y Financieros, Universidad Miguel Hernández, Spain

ARTICLE INFO

JEL classification:
D82
D83
K40

Keywords:
Multiple types
Career concerns
Transparency on consequences
Perverse effect

ABSTRACT

Quite often an expert takes position on an issue where certain actions can be perceived as biased. If the expert has an informational concern and she does not want the listener to perceive her as biased, she has an incentive to avoid the biased action, even if she thinks this is the correct action. This paper shows that when an expert has multiple types and two concerns, an informational concern and a bias concern, the incentive to contradict private relevant information and avoid the biased action can even increase when the listener *observes* the quality of the expert's advice. We provide necessary and sufficient conditions for this perverse effect of transparency on consequences to emerge and discuss variations of the model.

1. Introduction

Transparency is often thought to be fundamental to a well-functioning institution. When experts have valuable information to transmit to an uninformed listener, the fact they are accountable to the listener often makes the latter feel secure, as the listener can feel it is more likely that information flows and less likely that it does in a way that fools the listener. When experts can be biased towards some piece of information, the hope that transparency disciplines experts and alleviates the information transmission problem may even be greater. But, should a rational listener always expect transparency to have this positive effect?

Consider an elected politician taking position on the convenience of going one step further on the application of a policy, e.g. increasing the budget of an affirmative action policy that promotes representation of women and minorities in society. If the politician were sympathetic to organizations for the rights of minorities, she would increase funding under any circumstance. The voter would then rationally expect a proposal to increase funding to be a signal of the politician being biased, e.g. too close to feminist movements, which makes the latter reluctant to propose it. However, affirmative action policies play a role in society and devoting more resources to these programs can be appropriate under some circumstances. The politician, who cares about reelection and wants to be perceived as a wise and unbiased politician, doubts whether to increase funding or not. The voter thinks that this trade-off could resolve in favor of the correct policy if the voter could observe the politician's action and also its consequences, i.e., the effects of the policy on society. The drawback the voter does not realize is that with transparency on consequences the politician may be more prone to taking the safer unbiased action, as transparency can increase the burden of proof of a wrong biased policy. Unexpectedly, with more information the voter is worse off.

Taking position on affirmative action policies, illegal immigration measures, after-prison programs, and other such policies may not be easy. On the one hand, the relevance of these social issues sometimes makes people adopt ideological and sentimental

☆ We thank the Editor and two anonymous referees for useful comments. We gratefully acknowledge financial support from the Ministerio de Ciencia, Innovación y Universidades, Spain (MCIU/AEI/FEDER, UE) through projects RTI2018-097620-B-I00, PGC2018-097965-B-I00, PID2021-127736NB-I00, and PID2022-137211NB-I00, and the Junta de Andalucía, Spain through project P18-FR-3840. The usual disclaimer applies.

* Corresponding author.
E-mail addresses: aandina@uma.es (A. Andina-Díaz), jose.garciam@umh.es (J.A. García-Martínez).

https://doi.org/10.1016/j.ejpoleco.2023.102439
Received 11 October 2022; Received in revised form 10 May 2023; Accepted 6 July 2023
Available online 25 July 2023

positions, turning some experts (e.g. politicians or advisors) into behavioral experts—always supporting one course of action. The existence of behavioral experts can make non-behavioral ones care not only about the informativeness of their speech but also about them not being mistaken for the behavioral experts. On the other hand, the relevance of these issues can attract public attention turning private debate into public debate. Under the scrutiny of the media and the eyes of large audiences on experts' speeches, will experts make more accurate decisions?

This paper proposes a model that captures this situation. An expert (agent) has relevant information to communicate to a listener (principal). The agent is either a "biased" type – e.g. she always supports one course of action, irrespective of the state – or one of two other types: a perfectly informed type or an imperfectly informed type. The agent has a dual concern: a concern for being perceived as informed, as in career concern models, and a concern for being perceived as unbiased, as she dislikes being mistaken for the biased type. The agent takes the decision about the information to send aiming at maximizing this dual concern.

In this set-up, we find that the equilibrium behavior of unbiased agents depends on how numerous biased agents in the population are. In the extremes, when the likelihood the agent is biased is sufficiently low, unbiased agents have no fear and honestly transmit their information; whereas when this likelihood is high, the opposite occurs and all unbiased agents, even those perfectly informed, avoid the biased action and disregard informative signals (Proposition 3). These results do not vary with transparency on consequences, suggesting that whether the listener observes the quality of the expert's advice or not is here irrelevant. In contrast to this, when the likelihood that the agent is biased is "intermediate", transparency on consequences does affect agents' behavior, but possibly contrary to what is expected, as it can make experts contradict informative signals more often (Proposition 4). This result suggests that whether the listener observes the quality of the expert's advice or not is here an issue. In fact, more transparency can lead to greater bias and less accurate advice, suggesting that with multiple types, asymmetric bias, and two concerns, transparency on consequences can be a double-edged sword.

To understand this seemingly paradoxical result, we explore the forces behind the perverse effect of transparency on consequences and provide necessary and sufficient conditions for this result to hold (Remarks 1 and 2 and Proposition 5). We show that the result cannot occur unless the agent has the double concern. For a snapshot of the argument, suppose the agent only cares about being unbiased. This concern clearly pushes the agent towards the unbiased action. With increased transparency, this incentive increases as a mistaken biased action will be attributed to bias, whereas a mistaken unbiased action will not. The reason is that with transparency the principal treats the two actions asymmetrically.[1] He perceives as unbiased an agent that takes the unbiased action even if she is proven wrong, but he perceives as (likely to be) biased an agent that takes the biased action, even if she is proven right. This asymmetric burden of proof creates an incentive to take the unbiased action that increases with transparency.

As we will see, this sole effect is however not sufficient for the perverse effect to appear. The reason is that an agent with a sole concern for bias will never take the biased action. In short, the perverse effect is already there, though hidden. We need it to come to light, and here comes the informational concern. Because perfectly informed agents sometimes take the biased action, if the agent also cares about providing accurate advice, she will try to copy the informed type and will sometimes take the biased action. Hence, it is the combination of the two concerns that yields this perverse effect of transparency. To the best of our knowledge, this result and the mechanism behind it are new in the literature.

The rest of the paper is organized as follows. In Section 2 we review the related literature and in Section 3 we present the model. Section 4 contains the analysis and the main results, which are structured in three parts: preliminaries, main result, and forces behind the perverse effect. Section 5 discusses variations of the model and extends results to alternative frameworks. Finally, Section 6 concludes. All proofs are in Appendix.

2. Literature review

Our model is closely related to literature that studies the perverse effects of transparency on information revelation by career-concerned agents. The seminal paper in this literature is Prat (2005), which shows that transparency on actions can harm the principal. The crucial idea in this paper is the relative smartness of the realizations of the agent's signal, i.e., how similar or different the posteriors on the agent's type are for each realization of the agent's signal. Prat (2005) shows that when one realization is much smarter than the other, and the agent's action is informative of the signal, the agent has an incentive to take the action that corresponds to the smart realization of the signal. This incentive increases in transparency on actions but decreases in transparency on consequences. Hence, while transparency on consequences increases the principal's welfare, transparency on actions can harm the principal. Fox and Van Weelden (2012) show that transparency on consequences can be detrimental to the principal. The key assumption in this work is that the costs to the principal of the agent's mistakes are exogenously asymmetric across the states, in the sense that some mistakes are more costly than others. Under this set-up, the authors show that an increase in the probability that the principal learns the consequences of the agent's action may reduce the principal's expected welfare. Our model differs from these works in considering that states, signals, and error costs are symmetric. We consider instead an agent with multiple types and two concerns, which endogenously makes actions asymmetric. In common with these papers, we all play on an asymmetry, though the source of the asymmetry is a different one.

Within the electoral accountability literature, Canes-Wrone et al. (2001) propose a model that also plays on an asymmetry. They consider that different policies have different probabilities of transparency on consequences and show that when transparency is sufficiently asymmetric across actions, the politician faces an incentive to disregard informative private signals, which may reduce

[1] More precisely, with transparency the principal treats the two actions more asymmetrically than without transparency.

voters' welfare.[2] Devdariani and Hirsch (2022) extend the basic set-up in Canes-Wrone et al. (2001) and consider endogenous information acquisition by voters. They show that voter attention generally improves politicians' accountability, but it can also be harmful to the voters. The latter occurs when voter attention makes one policy much more attractive than the other, distorting the incumbent's policy.[3] Similar in spirit but considering exogenous information instead, Ashworth and Shotts (2010) build on Canes-Wrone et al. (2001) to include a media outlet. They show that a more informative media may exacerbate the incentive of the incumbent to pander to the ex-ante popular policy. We differ from these works in considering an agent with multiple types and pure career concerns, and do not rely on herding effects. Other recent papers looking at transparency on consequences are Blumenthal (2022)—which considers endogenous information acquisition – and Foerster and Voss (2022) – which consider an exogenous informative source and an agent with multiple types, like us. Both papers consider two-period models and show that voters might be worse off with more information, as transparency facilitates politicians' sorting but reduces the likelihood of imitation and, hence, of a better policy in the first period.[4] This argument is different from ours: In our case transparency can increase the likelihood of a pooling equilibrium, as transparency induces an agent to disregard an informative signal more often; whereas in their case transparency reduces the likelihood of a (good) pooling equilibrium.

Other papers analyzing transparency are Li and Madarász (2008) and Bourjade and Jullien (2011), which identify perverse effects of transparency on the agent's preferences; Holmström (1999) and Dewatripont et al. (1999), which show that transparency can induce agents to exert less effort; Levy (2007), Sibert (2003), and Gersbach and Hahn (2008), showing that more transparent mechanisms can lead to worse decisions in committees; Ashworth and de Mesquita (2014), showing that increases in voter information may make democratic performance worse; and Morris (2001), Ottaviani and Sørensen (2001), Hörner (2002), and Ely and Välimäki (2003), which identify perverse effects of reputation.

Finally, our paper is also related to literature considering agents with multi-dimensional types and two concerns. Austen-Smith and Fryer (2005), Bénabou and Tirole (2006), Esteban and Ray (2006), Bagwell (2007), and Frankel and Kartik (2019) analyze information transmission in the presence of an agent with two-dimensional types, and Austen-Smith and Fryer (2005), Fox and Shotts (2009), Bar-Isaac and Deb (2014), and Feller and Schäfer (2020) study information transmission in the presence of two audiences, i.e., two concerns. None of these papers analyze the effect of transparency on consequences on the quality of the decision-making process.

3. The model

We consider a model with a principal (he) and an agent (she). The agent has to take a binary action $a \in \{L, R\}$ to match the state of the world $\omega \in \{L, R\}$. We assume action L is the *correct* action in state L and action R is the correct one in state R. Each state occurs with equal probability.[5] Prior to taking an action, the agent receives a private signal $s \in \{L, R\}$ about the state. The agent is either a biased type who always takes one action, or is of one of two other types: an informed type who observes the state or one who observes an imperfect informative signal of quality $\gamma \in (\frac{1}{2}, 1)$ about the state. We denote by B the biased type, by W the type that observes the state (also referred to as the wise type), and by N the type that observes the signal (also referred to as the normal type).[6] The agent knows her type but the principal does not. Let α_B, α_W, and α_N denote the prior probability that the agent is the biased, wise and normal type, respectively, with $\alpha_t > 0$ for $t \in \{B, W, N\}$ and $\sum_t \alpha_t = 1$. Without loss of generality, we assume the biased type always takes action L and we refer to this action as the *biased action*; hence, R is the unbiased action. The focus of the paper is on the behavior of the strategic types $t \in \{W, N\}$.[7] We denote the strategy of an agent by $\sigma_t(s) \in [0, 1]$, describing the probability that an agent of type t takes the biased action L after signal s.

The principal observes the action of the agent. Additionally, with probability $\mu \in [0, 1]$, he observes the state of the world. We refer to μ as the probability that there is *transparency on consequences*. We denote by $X \in \{0, L, R\}$ the feedback received by the principal, with L (alternatively, R) indicating the principal learns that the state is L (alternatively, R), and 0 indicating that there is no feedback.

Upon observing the action $a \in \{L, R\}$ and feedback $X \in \{0, L, R\}$, the principal updates his belief about the type of the agent. Let $\lambda_B(a, X)$, $\lambda_W(a, X)$, and $\lambda_N(a, X)$ denote the principal's belief that the agent is a biased, wise, and normal type, respectively. The principal receives utility 1 when the action is the correct one given the state, i.e., $a = \omega$; and he receives utility 0 otherwise. Note that because the signal of the agent is informative, i.e., $\gamma > 1/2$, the principal's expected welfare is maximized when the agent follows her signal, i.e., $a = s$. Based on this, we will say that transparency on consequences is detrimental to the principal, i.e., it has a perverse effect, when it refrains the agent from following her signal and induces her to take action $a \neq s$.

[2] See Levy (2005), Leaver (2009), Liu and Sanyal (2012) and Andina-Díaz and García-Martínez (2020) for other papers with asymmetric monitoring.

[3] In their model, this occurs when voter attention makes a strong (weak) incumbent that would secure (lose) reelection without attention, distort his policy towards one that secures reelection by evading (drawing) voter attention.

[4] This occurs when, in equilibrium without information, different types of politicians pool in the "correct" action, for this action enhances the probability of reelection.

[5] This assumption guarantees that herding motives are not behind our results. We relax this assumption in Section 5 and show the robustness of our main results.

[6] This setup is analogous to one considering the agent has two dimensions of private information: ability (to be informed about the state) and bias (or preference for some action), provided that biased types always take one action.

[7] See Section 5 for an extension that considers a strategic biased type.

The agent has reputational concerns and cares about both her reputation for being informed and her reputation for being unbiased. For a given action a and feedback X, the payoff to the agent is:

$$\Pi(a,X)=\eta\lambda_W(a,X)+\beta\lambda_{\bar{B}}(a,X), \tag{1}$$

where $\lambda_{\bar{B}}(a,X)=1-\lambda_B(a,X)$ is the probability the agent is unbiased and $\eta,\beta>0$ represent the weights of the two concerns, being informed and being unbiased, respectively.[8] Since at the time of taking the action the agent only knows the probability μ that the principal will learn the state, for a given type t and signal s, the agent chooses action a that maximizes her expected payoff:

$$\Pi^{\mu}_{t,s}(a)=(1-\mu)\Pi(a,0)+\mu[P(\omega=L|s;t)\Pi(a,L)+P(\omega=R|s;t)\Pi(a,R)]. \tag{2}$$

Our equilibrium concept is Perfect Bayesian Equilibrium. We will denote an equilibrium strategy by $\sigma^{\mu*}=(\sigma^{\mu*}_W,\sigma^{\mu*}_N)$, with $\sigma^{\mu*}_W=(\sigma^{\mu}_W(L)^*,\sigma^{\mu}_W(R)^*)$ and $\sigma^{\mu*}_N=(\sigma^{\mu}_N(L)^*,\sigma^{\mu}_N(R)^*)$ for the wise and the normal type, respectively.

4. Results

For expositional purposes, the presentation of the results considers two simplifications. First, we skip the limit case of $\mu=0$ and focus instead on the general and more interesting case of $\mu>0$. Briefly, when $\mu=0$, there is a multiplicity of equilibria, which introduces many particularities in the results and a need for equilibrium selection. To facilitate the reading process, we relegate the results of the case $\mu=0$ to Appendix. Second, we restrict our analysis to non-perverse equilibria. We say that an equilibrium is *non-perverse* if for a given μ, $\sigma^{\mu}_t(L)^*\geq\sigma^{\mu}_t(R)^*$ for all $t\in\{W,N\}$, i.e., the two strategic types use non-perverse strategies.

Other concepts that we use are the following. For a given type $t\in\{W,N\}$, we say that an equilibrium strategy is *informative* if $\sigma^{\mu}_t(L)^*\neq\sigma^{\mu}_t(R)^*$, i.e., the strategy is signal dependent, and non-informative otherwise. An equilibrium will be informative when the two strategic types use informative strategies and non-informative otherwise. Additionally, for a given type t, an informative equilibrium strategy is *honest* if $(\sigma^{\mu}_t(L)^*,\sigma^{\mu}_t(R)^*)=(1,0)$.

4.1. Preliminaries

As we will see, the perverse effect of transparency is the result of a subtle combination of incentives that the agent can face. Walking the reader to this result requires some initial steps. This is the purpose of this section, which is structured as follows. First, we characterize the behavior of the wise type, which allows us to argue why the posterior analysis focuses on equilibria in which the wise type uses an honest strategy. Second, we characterize the behavior of the normal type after signal R and then argue why the rest of the paper focuses on the behavior of the normal type after the "biased" signal L. Finally, we analyze the equilibrium behavior of the normal type after signal L when the number of biased types is either sufficiently high or low. The analysis of the equilibrium behavior of the normal type after signal L when the number of biased types is "intermediate" will be the focus of Section 4.2.

The first result characterizes the behavior of a wise type. It distinguishes two cases, according to whether the strategy of the normal type is informative or not.

Proposition 1. *For any $\mu>0$:*

1. *If in equilibrium the normal type plays an informative strategy, i.e., $\sigma^{\mu}_N(L)^*\neq\sigma^{\mu}_N(R)^*$, then the unique equilibrium strategy of the wise type is the honest strategy, i.e., $(\sigma^{\mu}_W(L)^*,\sigma^{\mu}_W(R)^*)=(1,0)$.*
2. *Otherwise, i.e., $\sigma^{\mu}_N(L)^*=\sigma^{\mu}_N(R)^*$, there exists $\widetilde{\alpha}_B\in(0,1)$ such that:*
   (a) *If $\alpha_B>\widetilde{\alpha}_B$, then the unique equilibrium strategy of the wise type is to always take action R, i.e., $(\sigma^{\mu}_W(L)^*,\sigma^{\mu}_W(R)^*)=(0,0)$.*
   (b) *If $\alpha_B<\widetilde{\alpha}_B$, then in equilibrium the wise type always takes action R after signal R and, when the signal is L, takes action L with probability x_1, i.e., $(\sigma^{\mu}_W(L)^*,\sigma^{\mu}_W(R)^*)=(x_1,0)$ with $x_1\in[0,1]$.*

For expositional purposes, we first discuss point *2.* of the Proposition, which characterizes the behavior of the wise type when the normal type uses a non-informative strategy. It identifies the existence of threshold $\widetilde{\alpha}_B$ such that when the probability that the agent is biased is higher than the threshold, in equilibrium the wise type always plays a non-informative strategy that consists on taking the unbiased action R always, for any signal. When α_B is however lower than the threshold, in equilibrium the wise type can either play the same non-informative strategy $(\sigma^{\mu}_W(L)^*,\sigma^{\mu}_W(R)^*)=(0,0)$ or rather an informative strategy $(\sigma^{\mu}_W(L)^*,\sigma^{\mu}_W(R)^*)=(x_1,0)$, with $x_1>0$. This result suggests that a high α_B can even induce the wise type to avoid the biased action, for fear of being tarred as biased. We will come back to this idea later in the text.

The first point of Proposition 1 characterizes the behavior of the wise type when the normal type uses an informative strategy. It states that, in this case, the wise type always follows her signal in equilibrium. This result is quite intuitive, as provided that the

[8] There are two comments on this. First, some of our results hold for any payoff function $\Pi(a,X)=f(\lambda_W(a,X),\lambda_{\bar{B}}(a,X))$ that is increasing in both arguments. When applying, we specify it in the proof of the result in Appendix. Second, the payoff function in (1) can be easily microfoundated by considering that the principal assigns a reputational rent π_θ to each type $\theta\in\{B,W,N\}$, so the agent maximizes $\Pi(a,X)=\lambda_B(a,X)\pi_B+\lambda_W(a,X)\pi_W+\lambda_N(a,X)\pi_N$. Assuming $\pi_W>\pi_N>\pi_B$ and $\pi_B=0$, and using $\lambda_B(a,X)=1-\lambda_B(a,X)$, we get expression (1).

Table 1
Principal's belief that the agent is an informed type.

$\lambda_W(a,X)$			
	$X=0$	$X=L$	$X=R$
$a=L$	$\frac{\alpha_W}{2\alpha_B+\alpha_W+\sigma_N(L)\alpha_N}$	$\frac{\alpha_W}{\alpha_B+\alpha_W+\gamma\sigma_N(L)\alpha_N}$	0
$a=R$	$\frac{\alpha_W}{\alpha_W+(2-\sigma_N(L))\alpha_N}$	0	$\frac{\alpha_W}{\alpha_W+(\gamma+(1-\gamma)(1-\sigma_N(L)))\alpha_N}$

Table 2
Principal's belief that the agent is an unbiased type.

$\lambda_{\bar{B}}(a,X)$			
	$X=0$	$X=L$	$X=R$
$a=L$	$\frac{\alpha_W+\sigma_N(L)\alpha_N}{2\alpha_B+\alpha_W+\sigma_N(L)\alpha_N}$	$\frac{\alpha_W+\gamma\sigma_N(L)\alpha_N}{\alpha_B+\alpha_W+\gamma\sigma_N(L)\alpha_N}$	$\frac{(1-\gamma)\sigma_N(L)\alpha_N}{\alpha_B+(1-\gamma)\sigma_N(L)\alpha_N}$
$a=R$	1	1	1

normal type sometimes follows her signal, the wise type has a stronger incentive to follow hers—the signal of the latter is of higher quality.

A first implication of Proposition 1 is that condition $\sigma_N^\mu(L)^* \neq \sigma_N^\mu(R)^*$, i.e., the normal type uses an informative strategy, is necessary and sufficient for the equilibrium to be informative. A second implication is that in any informative equilibrium, the wise type uses the honest strategy. In other words, there is no informative equilibrium in which $(\sigma_W^\mu(L)^*, \sigma_W^\mu(R)^*) \neq (1,0)$. Corollary 1 below formally states (and elaborates a bit further) this idea, by describing the behavior of the normal type when the wise type is not honest.

Corollary 1. *For all $\mu > 0$, if in equilibrium the wise type is not honest, i.e., $(\sigma_W^\mu(L)^*, \sigma_W^\mu(R)^*) \neq (1,0)$, then the unique equilibrium strategy of the normal type is to always take action R, i.e., $(\sigma_N^\mu(L)^*, \sigma_N^\mu(R)^*) = (0,0)$.*

Corollary 1 states that if the wise type does not use the honest strategy, then, in equilibrium, the normal type always takes the unbiased action R, for any level of transparency. A straightforward implication of Proposition 1 and Corollary 1 is that there is always an equilibrium in which the two strategic types pool at the unbiased action R. In other words, the strategy profile $(\sigma_W^\mu(L)^*, \sigma_W^\mu(R)^*; \sigma_N^\mu(L)^*, \sigma_N^\mu(R)^*) = (0,0;0,0)$ is always an equilibrium strategy profile. Proposition 3 will further elaborate on this idea.

We next move on to the analysis of the normal type and first consider the behavior after signal R. Note that in this case there is no tradeoff, as action R is optimal in terms of the agent's desire to look both informed and unbiased. As a consequence, the agent never takes action L and the next result follows.

Proposition 2. *For all $\mu > 0$, the normal type always takes action R after signal R, i.e., $\sigma_N^\mu(R)^* = 0$.*

Then, the rest of the paper analyzes the behavior of the normal type after signal L. The reader might expect this behavior to change depending on how numerous biased types in the population are. In fact, it is the existence of biased types that makes action L be perceived as biased. We start the analysis with this exercise. To this aim, we consider $\sigma_N^\mu(R)^* = 0$ and $(\sigma_W^\mu(L)^*, \sigma_W^\mu(R)^*) = (1,0)$, and analyze the beliefs of the principal about the type of the agent.[9] For a given action $a \in \{L,R\}$ and feedback $X \in \{0,L,R\}$, Tables 1 and 2 below describe the principal's beliefs $\lambda_W(a,X)$ and $\lambda_{\bar{B}}(a,X)$:

Consider first that the prior probability that the agent is biased is very high, i.e., $\alpha_B \to 1$. In this case, we can observe that both $\lambda_W(L,X)$ and $\lambda_{\bar{B}}(L,X)$ tend to 0 for any $X \in \{0,L,R\}$ and, therefore, for any μ. Hence, the first idea that we draw is that when the prior probability that the agent is biased is sufficiently high, the optimal choice of a normal type who receives signal L is the unbiased action R, even if transparency is very high. Moreover, we can show that, in this case, there is no equilibrium in which $(\sigma_W^\mu(L)^*, \sigma_W^\mu(R)^*) = (1,0)$, and that in the unique equilibrium, $\sigma_t^\mu(L)^* = 0$ for all $t \in \{W,N\}$ and $\mu > 0$. This means the unique equilibrium in this case is the non-informative equilibrium where all strategic types take action R. This result, which constitutes the first point of Proposition 3 below, suggests that when the perception that the agent is biased is higher than a certain threshold, the incentive to avoid looking biased is so strong that, in the unique equilibrium of the game, agents disregard (even perfect) informative signals and always take the unbiased action—even if they only have an informational concern.[10]

Suppose now that the prior probability that the agent is biased is very low, i.e., $\alpha_B \to 0$. In this case, we can observe that $\lambda_{\bar{B}}(L,X) \to \lambda_{\bar{B}}(R,X) = 1$ for any $X \in \{0,L,R\}$ and, therefore, for any μ. This means that, in the limit, the principal's belief that an agent is unbiased is invariant to the agent's action and to the level of transparency. As a result, the normal type's decision is, here, exclusively driven by her informational concern, $\lambda_W(a,X)$. In this case, we obtain that there is an equilibrium in which the normal type always takes action L after signal L, for any level of transparency. This result constitutes the second point of Proposition 3.

[9] This is the interesting case as, from Corollary 1, if $(\sigma_W^\mu(L)^*, \sigma_W^\mu(R)^*) \neq (1,0)$, the normal type always takes the unbiased action, for any $\mu > 0$.

[10] A sufficient condition for $\sigma_N^\mu(L)^* = 0$ is $\alpha_B > 1/2$. See Lemma 7 in Appendix A.3.

Proposition 3. *For any $\mu > 0$, there exists $\alpha_B^{max} < 1$ and $\alpha_B^{min} > 0$ such that the following holds:*

1. *If $\alpha_B > \alpha_B^{max}$, the unique equilibrium is non-informative, with all types of agents taking action R, i.e, $(\sigma_W^\mu(L)^*, \sigma_W^\mu(R)^*) = (0,0)$ and $(\sigma_N^\mu(L)^*, \sigma_N^\mu(R)^*) = (0,0)$.*
2. *If $\alpha_B < \alpha_B^{min}$, the unique informative equilibrium is the honest equilibrium, i.e., $(\sigma_W^\mu(L)^*, \sigma_W^\mu(R)^*) = (1,0)$ and $(\sigma_N^\mu(L)^*, \sigma_N^\mu(R)^*) = (1,0)$.*

Two conclusions are drawn from Proposition 3. First, the consideration of a biased type – who always takes the biased action – has important effects on the behavior of the strategic types and, especially, on the behavior of the normal type.[11] In particular, we show that a sufficiently high α_B induces the normal type (and even the wise type) to always take the unbiased action, for fear of looking biased; whereas a sufficiently low α_B induces the normal type to always follow her signal, trying to look informed. Second, when the proportion of biased types in the population is either too high or too low, transparency on consequences does not affect the equilibrium behavior of the strategic types. This result suggests that the perverse effect, if appearing, requires "intermediate" values of α_B.

4.2. The perverse effect of transparency

This section presents the main result of the paper, stated in Proposition 4. It shows that there exists a region of parameters in which an increase in the probability that the principal learns the state of the world induces a normal type that receives signal L to disregard it more often and to take the unbiased action R with a higher probability. The conditions on the parameters refer to the prior probability that the agent is biased, which must be "intermediate", and the quality of the signal of the agent, which cannot be very high. The expressions of these thresholds are given in the proof of the result, in Appendix A.3. For simplicity, the result assumes $\eta = \beta$, so that a normal type cares equally about her reputation for unbiasedness and her reputation for being informed. However, this assumption is not crucial to the result, as Proposition 5 in the next section shows.

Proposition 4. *Consider $\eta = \beta$. There exist cutoffs γ', μ', μ'', α_B', and α_B'', with $\gamma' > \frac{1}{2}$, $0 < \mu' < \mu'' < 1$, and $0 < \alpha_B' < \alpha_B'' < \frac{1}{2}$, such that for all $\gamma \in (\frac{1}{2}, \gamma')$ and $\alpha_B \in (\alpha_B', \alpha_B'')$:*

1. *If $\mu < \mu'$, in the unique informative equilibrium the wise type is honest, i.e., $(\sigma_W^\mu(L)^*, \sigma_W^\mu(R)^*) = (1,0)$, and the normal type is honest after signal R and, after signal L, is honest with probability σ', i.e., $(\sigma_N^\mu(L)^*, \sigma_R^\mu(R)^*) = (\sigma', 0)$, with $0 < \sigma' < 1 - \frac{2\alpha_B\alpha_W + \alpha_B\alpha_N}{(\alpha_W - \alpha_B)\alpha_N}$ and σ' decreasing in μ.*
2. *If $\mu \in (\mu', \mu'')$, in the most informative equilibrium the wise type is honest, i.e., $(\sigma_W^\mu(L)^*, \sigma_W^\mu(R)^*) = (1,0)$, and the normal type is honest after signal R and, after signal L, is honest with probability σ'', i.e., $(\sigma_N^\mu(L)^*, \sigma_N^\mu(R)^*) = (\sigma'', 0)$, with $0 < \sigma'' < \sigma'$ and σ'' decreasing in μ.*
3. *If $\mu > \mu''$, there is no informative equilibrium. In this case, the normal type always takes action R, i.e, $(\sigma_N^\mu(L)^*, \sigma_N^\mu(R)^*) = (0,0)$ and the wise type plays $(\sigma_W^\mu(L)^*, \sigma_W^\mu(R)^*) = (x_1, 0)$, with $x_1 \in [0,1]$.*

Proposition 4 describes the equilibrium behavior of the strategic types for different degrees of transparency on consequences.[12] The first point of the proposition states that for low enough values of μ, there is a unique informative equilibrium, in which the normal type follows signal L with probability σ', with σ' decreasing in μ.[13] The second point considers intermediate values of μ, in which case there can be multiple equilibria.[14] Here we focus on the most informative equilibrium, which is the most interesting equilibrium from the point of view of the principal—it is the one that maximizes his welfare. We obtain that the equilibrium probability that the normal type follows signal L, σ'', also decreases in μ, with $\sigma'' < \sigma'$. The third point states that for high enough values of μ, the normal type always takes the unbiased action R; which implies information transmission completely breaks down.

[11] The consideration of $\alpha_B > 0$ has a first effect on the equilibrium behavior of the normal type. The proof of point *1.* of Proposition 3 shows that for α_B higher than a certain threshold, the unique equilibrium strategy of the normal type is non-informative. Additional increases in α_B can also affect the equilibrium behavior of the wise type. In particular, we show that for α_B that is high enough, in equilibrium, the wise type also uses a non-informative strategy.

[12] In the proof of Proposition 4 (see Lemma 10), in Appendix A.3, we show that $\alpha_B'' < \widetilde{\alpha}_B$. Note that, by Proposition 1, if $\alpha_B < \widetilde{\alpha}_B$, then there is an equilibrium in which $(\sigma_W^\mu(L)^*, \sigma_W^\mu(R)^*) = (1,0)$.

[13] In Appendix A.3, we show that when $\mu = 0$, $\sigma_N^0(L)^* = 1 - \frac{2\alpha_B\alpha_W + \alpha_B\alpha_N}{(\alpha_W - \alpha_B)\alpha_N}$. Thus, $\sigma' < \sigma_N^0(L)^*$ for all $\mu > 0$.

[14] To explain the multiplicity, we use Definition 1 introduced in Section 4.3, and note that $\Delta_{N,L}^\mu$ is not necessarily a monotonic function in $\sigma_N^\mu(L)$. See Fig. 1. To see why $\Delta_{N,L}^\mu$ is not monotonic, first note that when $\sigma_N(L)$ increases, there are two forces at work: (i) the principal's belief that the agent is wise when taking action R increases (or does not vary) and the principal's belief that the agent is wise when taking action L decreases (or does not vary). See Table 1. Thus, the expected payoff to the agent for taking action R (L) increases (decreases); hence, $\Delta_{N,L}^\mu = \Pi_{N,L}^\mu(R) - \Pi_{N,L}^\mu(L)$ increases. See expression (4) in Appendix A.1. (ii) The principal's belief that the agent is unbiased when taking action R does not vary, and the principal's belief that the principal is unbiased when taking action L increases. See Table 2. Thus, the expected payoff to the agent for taking action L (R) increases (does not vary); hence, $\Delta_{N,L}^\mu = \Pi_{N,L}^\mu(R) - \Pi_{N,L}^\mu(L)$ decreases. See expression (4) in Appendix A.1. Second, note that when $\sigma_N(L)$ is low, effect (ii) is stronger than effect (i); hence, $\Delta_{N,L}^\mu$ decreases in $\sigma_N(L)$. In contrast, when $\sigma_N(L)$ is high, effect (i) is stronger than effect (ii); hence, $\Delta_{N,L}^\mu$ increases in $\sigma_N(L)$.

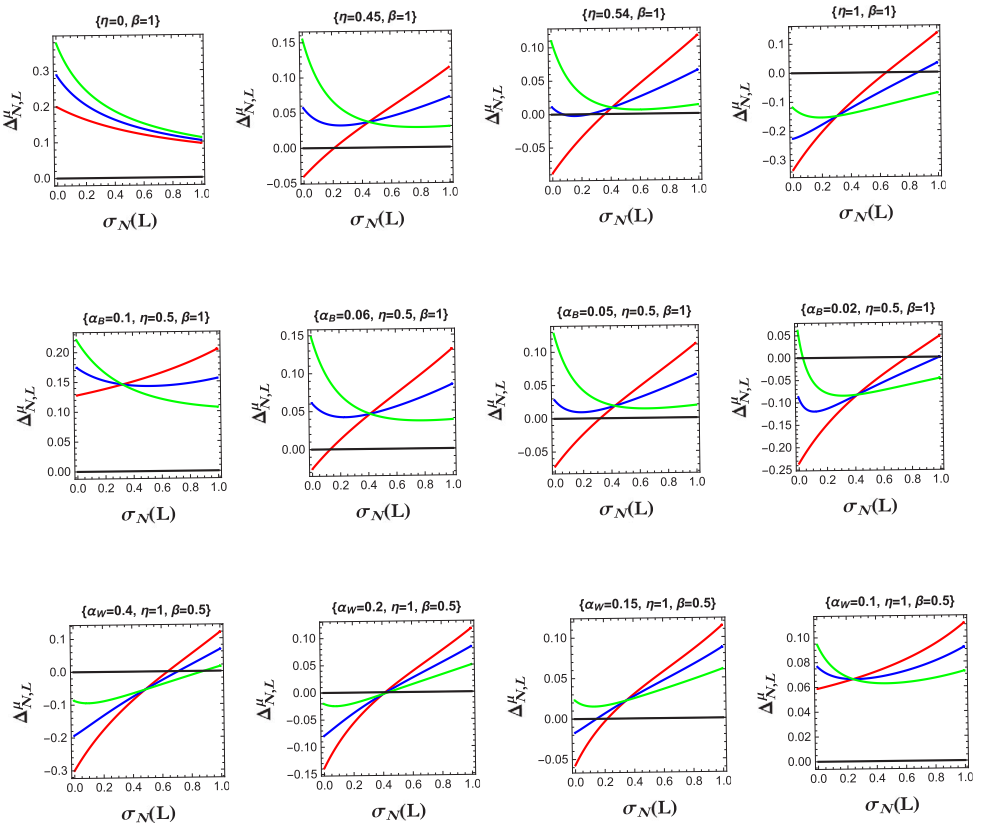

Fig. 1. We represent $\Delta^{\mu}_{N,L}$ for different degrees of transparency on consequences and different values of parameters η, α_B and α_W. Red functions represent $\Delta^{0}_{N,L}$, blue functions represent $\Delta^{0.5}_{N,L}$, and green functions represent $\Delta^{1}_{N,L}$. Upper panels represent variations in η, with $\eta = 0, 0.45, 0.54, 1$, from left to right; setting $\beta = 1$, $\alpha_W = 0.4$, $\alpha_B = 0.05$, and $\gamma = 0.7$. Middle panels represent variations in α_B, with $\alpha_B = 0.1, 0.06, 0.05, 0.02$, from left to right; setting $\eta = 0.5$, $\beta = 1$, $\alpha_W = 0.4$ and $\gamma = 0.7$. Bottom panels represent variations in α_W, with $\alpha_W = 0.4, 0.2, 0.15, 0.1$, from left to right; setting $\eta = 1$, $\beta = 0.5$, $\alpha_B = 0.1$ and $\gamma = 0.6$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.3. The sources of the perverse effect

At this point, it is clear there is in the model an incentive to take the unbiased action R but, why does this incentive increase with transparency on consequences? To answer this question we look into the driving forces behind the result, which are clarified when we separate the two concerns. For this reason, in this section we consider each of the two concerns separately and, for each of them, analyze the effect that an increase in μ has on the incentive of the normal type to contradict signal L.

The bias concern. Suppose the agent has a sole concern for being perceived as unbiased. This corresponds to the case $\eta \to 0$, hence the agent's payoff in (1) is simply $\Pi(a, X) = \beta\lambda_{\bar{B}}(a, X)$. In this case, we obtain that when the proportion of biased types in the population is below a certain threshold, the incentive of the normal type to contradict signal L is higher with transparency than without it. We formulate this result in Remark 1 below and before define the concept of expected gain.[15]

Definition 1. Let $\Delta^{\mu}_{t,s}$ be the expected gain to an agent of type t for taking action R rather than L after signal s when the level of transparency is μ, i.e., $\Delta^{\mu}_{t,s} = \Pi^{\mu}_{t,s}(R) - \Pi^{\mu}_{t,s}(L)$.

Remark 1. Consider $\eta \to 0$. Then, $\Delta^{0}_{N,L} < \Delta^{1}_{N,L}$ if and only if $\alpha_B < \frac{1-\gamma}{2\gamma-1}\alpha_W$.

To give an intuition for this result,[16] note that transparency on consequences has very different effects on the payoffs of the agent: While the payoff to the normal type from taking action R does not depend on μ, her payoff from taking the biased action L may

[15] The expected gain $\Delta^{\mu}_{t,s}$ is a function of the strategy profile (σ_W, σ_N), with $\sigma_W = (\sigma_W(L), \sigma_W(R))$ and $\sigma_N = (\sigma_N(L), \sigma_N(R))$. For the sake of simplicity, this dependence is sometimes omitted.

[16] The proof is straightforward. Simply note that when $\eta \to 0$, $\Delta^{0}_{N,L} = \frac{2\alpha_B}{2\alpha_B+\alpha_W+\sigma_N(L)\alpha_N}$ and $\Delta^{1}_{N,L} = \frac{\gamma\alpha_B}{\alpha_B+\alpha_W+\gamma\sigma_N(L)\alpha_N} + \frac{(1-\gamma)\alpha_B}{\alpha_B+(1-\gamma)\sigma_N(L)\alpha_N}$ (see Table 2).

decrease with μ. The reason for this asymmetry is that a type that takes action L and proves wrong – which requires transparency – reveals she is not wise and so signals she is biased with a higher probability. Without transparency, however, action L is not such a strong signal of bias, as wise types also take this action. In contrast to this, action R is always a signal of unbiasedness, and transparency neither increases nor decreases its strength. Consequently, transparency on consequences increases the incentive of the agent to take the unbiased action, for transparency increases the asymmetric burden of proof of the two actions. Furthermore, the higher the probability and/or the cost of mismatching the state, the higher the incentive to take the unbiased action.

This argument helps explain why condition $\alpha_B < \frac{1-\gamma}{2\gamma-1}\alpha_W$ is easier to satisfy the higher α_W and the smaller α_B. To see it, note that the higher α_W, the higher the cost of mismatching the state, as the higher it is the increase in the probability of being biased when being shown wrong. Regarding α_B, a similar idea applies, as when α_B is very high, the cost of mismatching the state is very low. Last, condition $\alpha_B < \frac{1-\gamma}{2\gamma-1}\alpha_W$ is also easier to satisfy the smaller γ, as the smaller γ, the higher the probability of mismatching the state and being shown wrong when taking action L; hence, the higher the incentive to take action R.

Last, note that the result of Remark 1, though crucial, does not suffice to explain the perverse effect of transparency on consequences. This is clear from the fact that when $\eta \to 0$, in equilibrium the normal type always takes action R, for all μ (see Table 2). The reason is that when there is a sole concern for bias, the incentive to take the unbiased action is so strong that it completely hides the asymmetric effect of transparency. In this sense, Remark 1 describes a (first) necessary condition, although not sufficient, for the perverse effect to appear.

The informational concern. Suppose now the agent has a sole concern for being perceived as informed. This corresponds to the case $\beta \to 0$, hence the agent's payoff in (1) is simply $\Pi(a, X) = \eta\lambda_W(a, X)$. Quite intuitively, adding an informational concern (on top of a bias concern) reduces the incentive to take the unbiased action and, eventually, allows the perverse effect of transparency to come to light. However, for this to occur, the following condition must hold.[17]

Remark 2. Consider $\beta \to 0$. Then, $\sigma_N^0(L)^* > 0$ if and only if $\alpha_B < \frac{1-\alpha_W}{2}$.

This result says that when the proportion of biased types is below a certain threshold, then, in equilibrium without transparency, the normal type follows signal L with positive probability. This remark describes a (second) necessary condition for the perverse effect to appear, i.e, $\alpha_B < \frac{1-\alpha_W}{2}$. In fact, if the agent were rather to always take action R when $\mu = 0$, i.e., $\sigma_N^0(L)^* = 0$, transparency would never produce a perverse effect, for it never increases the probability the agent contradicts signal L.

Note also that condition $\alpha_B < \frac{1-\alpha_W}{2}$ (which is equivalent to $\alpha_B < \alpha_N$) is easier to satisfy the smaller both α_W and α_B (alternatively, the higher α_N). To see this, note that the smaller α_B, the smaller the incentive to take action R. On the other hand, the smaller α_W, the more likely action R comes from a normal type, which also reduces the incentive to take action R.

In addition to this result, and in line with what the reader might expect, Proposition 9 in Appendix A.3 shows that when there is a sole informational concern, in the informative equilibrium transparency on consequences never decreases the probability that the normal type follows signal L. Because transparency disciplines in this case, when adding an informational concern (on top of a bias concern), we must be careful that the positive effect of transparency (coming from the informational concern) does not offset the negative effect of transparency (coming from the bias concern). The next result considers $\beta = 1$ and establishes sufficient conditions on η and the other parameters of the model for the perverse effect of transparency to come to light. The expressions of these thresholds are given in the proof of the result, in Appendix A.3.

Proposition 5. *Consider $\beta = 1$. There exists cutoffs $\bar{\alpha}_B$, $\bar{\gamma}$, $\bar{\mu}'$, $\bar{\mu}''$, $\bar{\eta}'$, and $\bar{\eta}''$, with $\bar{\alpha}_B \in (0,1)$, $\bar{\gamma} \in (\frac{1}{2},1)$, $0 < \bar{\mu}' < \bar{\mu}'' < 1$, and $0 < \bar{\eta}' < \bar{\eta}''$, such that for all $\alpha_B < \bar{\alpha}_B$, $\gamma < \bar{\gamma}$ and $\eta \in (\bar{\eta}', \bar{\eta}'')$:*

1. *If $\mu < \bar{\mu}'$, in the unique informative equilibrium $(\sigma_W^\mu(L)^*, \sigma_W^\mu(R)^*) = (1,0)$ and $(\sigma_N^\mu(L)^*, \sigma_N^\mu(R)^*) = (\bar{\sigma}', 0)$, with $\bar{\sigma}' > 0$.*
2. *If $\mu > \bar{\mu}''$, there is no informative equilibrium. In this case, $(\sigma_N^\mu(L)^*, \sigma_N^\mu(R)^*) = (0,0)$ and $(\sigma_W^\mu(L)^*, \sigma_W^\mu(R)^*) = (x_1, 0)$, with $x_1 \in [0,1]$.*

As expected, we observe that for the perverse effect of transparency to come to light, the weight of the informational concern, measured by η, cannot be too high nor too small. We also observe that the smaller γ and α_B, the higher the range of values for which the perverse effect appears. These effects are in line with the previous discussion of parameters.[18] Two further comments on the role of these parameters are worth mentioning. First, despite the perverse effect is easier to sustain the smaller γ, we can find parameter configurations for which the perverse effect exists for γ very high.[19] Second, for the perverse effect to appear, we do not need a large proportion of biased and wise types in the population. However, these groups could be very small (e.g., the example in footnote 19). This result suggests that what matters for our result is that these groups exist, so that there is a reason for having both an informational concern and a bias concern; and not that the size of these groups is large with respect to the whole population.

[17] To prove the result, first note that when $\beta \to 0$, $\Delta_{N,L}^0 = \frac{\alpha_W}{\alpha_W + (2-\sigma_N(L))\alpha_N} - \frac{\alpha_W}{2\alpha_B + \alpha_W + \sigma_N(L)\alpha_N}$ (see Table 1), with $\Delta_{N,L}^0$ being increasing in $\sigma_N^0(L)$ and $\Delta_{N,L}^0\big|_{\sigma_N^0(L)=1} > 0$. Additionally, note that $\Delta_{N,L}^0\big|_{\sigma_N^0(L)=0} = \frac{\alpha_W}{\alpha_W+2\alpha_N} - \frac{\alpha_W}{2\alpha_B+\alpha_W}$, with $\Delta_{N,L}^0\big|_{\sigma_N^0(L)=0} < 0$ iff $\alpha_B < \frac{1-\alpha_W}{2}$. From here, the result follows.

[18] The effect of α_W, which is implicit in some of the expressions of the thresholds above, can neither be low (from Remark 1) nor high (from Remark 2).

[19] For example, if $\gamma = 0.9$, $\eta = 0.000143$, $\beta = 1$, $\alpha_B = 0.00001$, and $\alpha_W = 0.175$, we obtain $\sigma_N^0(L)^* = 0.2$ and $\sigma_N^1(L)^* = 0$. This result suggests that even when the signal of the agent is very informative ($\gamma = 0.9$), an increase in transparency on consequences may induce the agent to contradict the signal more often..

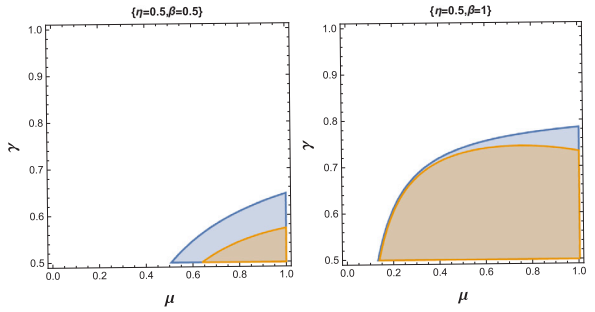

Fig. 2. In blue we represent the region of parameters (μ, γ) where the non-informative equilibrium $(\sigma_N^{\mu}(L)^*, \sigma_N^{\mu}(R)^*) = (0, 0)$ exists and in orange the region of parameters where this is the unique equilibrium. The panels consider $\eta = 0.5$, $\alpha_W = 0.4$, and $\alpha_B = 0.05$; with $\beta = 0.5$ in the left panel and $\beta = 1$ in the right panel. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Top panels of Fig. 1 illustrate the result of Proposition 5 with a numerical example. Given parameters α_B, α_W and γ, we observe that starting from $\eta = 0, \beta = 1$ and increasing η, reduces the incentive to contradict the signal and eventually makes the perverse effect of transparency come to light. To see it, note that both when $\eta = 0.45$ and $\eta = 0.54$, $\Delta_{N,L}^0$ (in red) is increasing and crosses the horizontal line at $\sigma_N^0(L) \in (0, 1)$. Note also that $\Delta_{N,L}^{0.5}$ (in blue) is either always positive (when $\eta = 0.45$) or it crosses the horizontal line at values $\sigma_N^{0.5}(L) < \sigma_N^0(L)$. Finally, note that $\Delta_{N,L}^1$ (in green) is always positive. From Definition 2 in the Appendix, this implies $\sigma_N^0(L)^* > \sigma_N^{0.5}(L)^* \geq \sigma_N^1(L)^* = 0$, i.e, honest behavior decreases with transparency. We also observe that increasing η too much results in transparency having the positive desirable effect, as the informational concern outweighs the bias concern. In fact, when $\eta = 1$, $\sigma_N^0(L)^* < \sigma_N^{0.5}(L)^* < \sigma_N^1(L)^* = 1$ in equilibrium, which implies honest behavior increases with transparency.[20]

Middle and bottom panels of Fig. 1 propose a different exercise. They play with parameters α_B and α_W. To understand the purpose of this exercise, note that both α_B and β affect the behavior of the agent in a similar way, as an increase in either parameter increases the incentive to contradict signal L—the bias concern becomes relatively more important. Similarly, an increase in either α_W or η plays also similar roles, as both make the informational concern relatively more salient. The substitutability between β and α_B on the one hand, and η and α_W on the other hand, although imperfect, has one important implication: it allows us to compensate effects and sustain the perverse effect of transparency on consequences even for quite different values of η and β. This is what middle and bottom panels of Fig. 1 show.[21] Middle panels propose a situation where the bias concern is twice stronger than the informational concern, i.e., $\eta = 0.5$ and $\beta = 1$, and vary α_B. We observe that despite the strong concern for bias, the perverse effect may come to light provided that the number of biased types is sufficiently small, for it reduces the incentive to contradict signal L. Bottom panels propose the opposite exercise. They consider a situation where the informational concern is twice stronger than the bias concern, i.e., $\eta = 1$ and $\beta = 0.5$, and vary α_W. Similarly to the previous case, we observe that the perverse effect may come to light provided that the number of wise types is sufficiently small, for it reduces the incentive to follow signal L.

Finally, Fig. 2 represents the region of parameters (μ, γ) for which the perverse effect of transparency can exist for a particular example. We consider parameters $\eta = 0.5$, $\alpha_W = 0.4$, and $\alpha_B = 0.05$; with $\beta = 0.5, 1$, in the left and the right panel, respectively. We represent in blue the region where the non-informative equilibrium $(\sigma_N^{\mu}(L)^*, \sigma_N^{\mu}(R)^*) = (0, 0)$ exists and in orange the region where this is the unique equilibrium (among informative and non-informative).[22] Note that since expressions (4) and (5) are continuous, an equilibrium always exists. Hence, the equilibria in the white area of the Cartesian plane are informative. We observe that the non-informative equilibrium is the unique equilibrium for μ sufficiently high and γ not too high; hence, Fig. 2 depicts the region where transparency has a perverse effect.

5. Discussion

This section discusses some of the assumptions of the model and shows the robustness of our results to these variations.

[20] A similar exercise can be done for β, starting from $\eta = 1, \beta = 0$, and increasing β. It is straightforward to see that an increase in β will increase the incentive to contradict signal L and, eventually, will make the perverse effect of transparency appears. For β sufficiently high, the incentive to contradict the signal will be so high that the agent will always take the unbiased action, irrespective of the level of transparency. This will occur when the bias concern outweighs the informational concern.

[21] Another idea is the relationship between parameters α_B and γ. Remember that when $\eta \to 0$, $\Delta_{N,L}^0 < \Delta_{N,L}^1$ if and only if $\alpha_B < \frac{1-\gamma}{2\gamma-1}\alpha_W$, which can be rewritten as $\gamma < \frac{\alpha_B+\alpha_W}{2\alpha_B+\alpha_W}$. Note that the expression on the right-hand side of the inequality is decreasing in α_B. Hence, the lower α_B is, the higher the range of values of γ for which the perverse effect of transparency on consequences appears.

[22] From Proposition 2, the profile $(\sigma_N^{\mu}(L)^*, \sigma_N^{\mu}(R)^*) = (0, 0)$ is always the unique non-informative equilibrium.

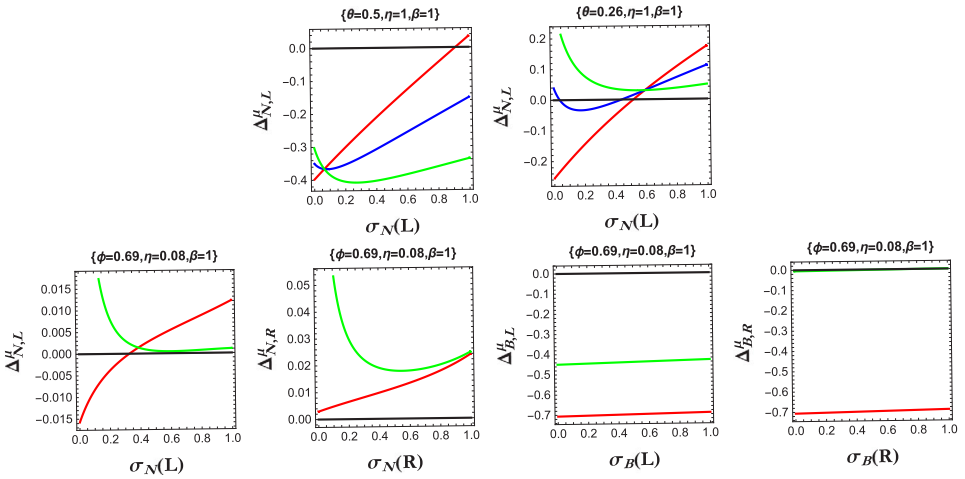

Fig. 3. Top panels represent $\Delta^{\mu}_{N,L}$ for different degrees of transparency on consequences and different values of parameter θ. Bottom panels represent $\Delta^{\mu}_{N,L}$, $\Delta^{\mu}_{N,R}$, $\Delta^{\mu}_{B,L}$, and $\Delta^{\mu}_{B,R}$, from left to right. The degrees of transparency of consequences we represent are $\mu = 0, 0.5, 1$ in top panels and $\mu = 0, 1$ in bottom panels. Red functions represent $\Delta^{0}_{t,s}$, blue functions represent $\Delta^{0.5}_{t,s}$, and green functions represent $\Delta^{1}_{t,s}$, for $t \in \{N, B\}$ and $s \in \{L, R\}$. Top panels consider $\eta = 1$, $\beta = 1$, $\gamma = 0.75$, $\alpha_W = 0.7$, and $\alpha_B = 0.01$; with $\theta = 0.5$ in the left panel and $\theta = 0.26$ in the right panel. Bottom panels consider $\phi = 0.69$, $\eta = 0.08$, $\beta = 1$, $\gamma = 0.7$, $\alpha_W = 0.2$, and $\alpha_B = 0.006$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.1. Popular belief about the state of the world

Here, we relax the assumption about the two states of the world being equally likely. Let θ be the prior probability that the state is L, with $\theta \in (0, 1)$. Note that considering one state to be more likely than the other introduces an incentive for the agent to go for the popular belief, as in herding models (see Avery and Chevalier (1999), Ottaviani and Sørensen (2006), and Gentzkow and Shapiro (2006), among others). Quite intuitively, this incentive to herd (on the popular belief) will reinforce or counterbalance the agent's incentive to take the unbiased action, depending on what the popular belief is. In other words, it may help the perverse effect come to light, or hide it instead, depending on the strength of the other forces at work—just in line with the aforementioned effects that the other parameters of the model, η, β, α_B, and α_W, have.

Top panels of Fig. 3 below present an exercise in this line. We consider $\eta = 1$, $\beta = 1$, and $\gamma = 0.75$, and propose a situation in which the proportion of wise types is so high as compared to the proportion of biased types ($\alpha_W = 0.7$ versus $\alpha_B = 0.01$), that the incentive to follow signal L outweighs the incentive to contradict it. As a result, in equilibrium, when $\theta = 1/2$, transparency has a positive desirable effect. Decreasing θ, however, increases the incentive to take the unbiased action R; hence, the incentive to go against the signal. This moves functions upward and, eventually, brings the perverse effect of transparency on consequences to light—as we observe in the right panel, with $\theta = 0.26$. A similar exercise can be done for the opposite case.[23]

The logic behind this exercise and the continuity of the agent's payoff function further suggest that if for a particular θ, say $\hat{\theta}$, the perverse effect appears in equilibrium, it would also appear for values of θ sufficiently close to $\hat{\theta}$. For the case of $\hat{\theta} = 1/2$, this argument suggests the robustness of our results to the consideration of a popular state of the world. The next result formally states this idea.

Proposition 6. *If $\sigma^1_N(L)^* < \sigma^0_N(L)^*$ when $\theta = 1/2$, there exists θ', θ'', with $\theta' < 1/2 < \theta''$, such that for all $\theta \in (\theta', \theta'')$, $\sigma^1_N(L)^* < \sigma^0_N(L)^*$.*

5.2. Strategic biased type

In the main body of the paper, we take the simplifying approach of considering that the biased type is not strategic and does not care about reputation. In this section, we relax this assumption. A simple way of analyzing a strategic biased type is to consider that this type maximizes a combination of career concerns and ideological concerns, say $\Pi(a, X) + \phi I$, with $\Pi(a, X)$ representing the career concerns, as described by (1), and ϕI representing the ideological concerns, with $I = 1$ if $a = L$ and $I = 0$ otherwise, and $\phi > 0$ describing the strength of this concern.[24] Apart from this, everything else in the model remains the same.

[23] We could describe a status quo scenario with $\theta = 1/2$, where the proportion of biased types is high as compared to that of wise types, so that the incentive to go against signal L outweighs the incentive to follow it. In this case, we expect the perverse effect to originate but not to appear in equilibrium when $\theta = 1/2$; however, an increase in θ, which increases the incentive to take the unbiased action, might eventually make the perverse effect come to light.

[24] There are two comments here. First, see Maskin and Tirole (2004) and Besley (2006), for papers where career concerned agents receive a benefit from taking *themselves* their preferred action. Second, an alternative approach would be to consider that the biased type aims to influence the action of the principal and persuades him to take her preferred action, as in Morris (2001). Though more complex than the alternative we analyze, we conjecture that for the appropriate discount factor (i.e., a sufficiently patient biased type), it would it an equilibrium in which the strategic biased type would always take the biased action.

Quite intuitively, a biased agent with strong enough ideological concerns would always take the biased action, which suffices to guarantee all our previous results hold. Next proposition states the result.

Proposition 7. *For all $\mu > 0$, there exits $\hat{\phi} > 0$ such that for all $\phi > \hat{\phi}$, $(\sigma_B^{\mu}(L)^*, \sigma_B^{\mu}(R)^*) = (1, 1)$.*

When the weight of the ideological concern goes below $\hat{\phi}$, the biased type does not always take action L but sometimes goes for the unbiased action R. Bottom panels of Fig. 3 propose an exercise in this line. We consider parameters $\phi = 0.69$, $\eta = 0.08$, $\beta = 1$, $\gamma = 0.7$, $\alpha_W = 0.2$, and $\alpha_B = 0.006$ and graphically show that the strategy profiles $(\sigma_W^0(L)^*, \sigma_W^0(R)^*; \sigma_N^0(L)^*, \sigma_N^0(R)^*; \sigma_B^0(L)^*, \sigma_B^0(R)^*) = (1, 0; 0.33, 0; 1, 1)$, and $(\sigma_W^1(L)^*, \sigma_W^1(R)^*; \sigma_N^1(L)^*, \sigma_N^1(R)^*; \sigma_B^1(L)^*, \sigma_B^1(R)^*) = (1, 0; 0, 0; 1, 0.8)$ are equilibria.[25] We observe that the biased type takes the unbiased action R in the equilibrium with transparency and transparency still has a perverse effect on the behavior of the normal type, as it induces her to go for action R more often. We also observe that transparency has a positive effect on the behavior of the biased type, as it induces this type to follow signal R more often. Altogether, however, the effect of transparency on consequences on the principal's welfare is detrimental, as the negative effect on the behavior of the normal type outweighs the positive effect on the behavior of the biased type—note that the proportion of biased types in the example is very low in comparison to the proportion of normal types, $\alpha_N \sim 0.8$ versus $\alpha_B = 0.006$. This result suggests that the perverse effect of transparency on the principal's welfare is robust to the consideration of a strategic biased type, provided that the probability this type takes the biased action is sufficiently high.

6. Conclusion

This paper analyzes the effect of transparency on consequences on the incentives of an expert (agent) to truthfully communicate her information to an uninformed listener (principal). We consider an "extended" career concerns model where the agent has multiple types, one of which is biased towards one policy, and two concerns: an informational concern and a bias concern. The introduction of the biased type and the concern for bias produces some new interesting insights.

There are three main insights from the model. First, the listener's perception of a bias in the population of experts has important consequences. Our results suggest that when there is a strong perception of biased expertise, unbiased experts overreact by taking the unbiased action too often. The fear of being tarred as biased can be so strong that it can even induce an expert with perfect information to contradict it, even if she only has an informational concern. Second, when the perception of biased expertise is either very high or low, transparency on consequences does not affect experts' behavior; their behavior is exclusively driven by their concern (either for being informed, unbiased, or both) and the relative number of biased experts in the population. Finally, when the number of biased experts is "intermediate", transparency on consequences does affect experts' behavior, but possibly contrary to what is expected, as it can make experts contradict informative signals more often. This result suggests that when there are multiple types and two concerns, transparency on consequences is a double-edged sword: It can be innocuous, it can be good, but it can also be harmful to the principal.

Our model helps rationalize the behavior of career concerned experts in the presence of a bias—e.g. gender or racial bias. It applies to experts in different contexts, from experts in the political arena and the judiciary to financial forecasters and health consultants. In all these contexts, our results have important policy implications. First, they suggest that seemingly benign interventions aimed at improving the availability of information might end up harming citizens. Second, interventions can be designed to avoid the perverse effect of transparency. This positive insight comes from the substitutability between some of the parameters of the model. It suggests that perverse incentives driven by extreme situations – e.g. from having a large number of biased types in a society – can be appropriately counterbalanced and offset, e.g. by introducing an "informational premium" that makes the informational concern more salient.

Declaration of competing interest

None.

Data availability

No data was used for the research described in the article.

[25] To understand the graphical analysis, from left to right panels, note that in the first panel, $\Delta_{N,L}^0$ (in red) is increasing and crosses the horizontal line at $\sigma_N(L) = 0.33$ and $\Delta_{N,L}^1$ (in green) is always positive. From Definition 2 in the Appendix, it implies $\sigma_N^0(L)^* = 0.33$ and $\sigma_N^1(L)^* = 0$. Similarly, in the second panel, note that both $\Delta_{N,R}^0$ and $\Delta_{N,R}^1$ are always positive; hence, $\sigma_N^0(R)^* = \sigma_N^1(R)^* = 0$. In the third panel, $\Delta_{B,L}^{\mu}$ is always negative, for all $\mu = \{0, 1\}$; hence, $\sigma_B^{\mu}(L)^* = 1$ $\forall \mu$. Finally, in the fourth panel, $\Delta_{B,R}^0$ is always negative and $\Delta_{B,R}^1$ is increasing and crosses the horizontal line at $\sigma_B(R) = 0.8$. It implies $\sigma_B^0(R)^* = 1$ and $\sigma_B^1(R)^* = 0.8$. Regarding the wise type, it can be shown that $\Delta_{W,L}^{\mu}$ and $\Delta_{W,R}$, evaluated at the strategy profiles in the text, are negative and positive, respectively. It implies $\sigma_W^{\mu}(L)^* = 1$ and $\sigma_W^{\mu}(R)^* = 0$ for all $\mu \in \{0, 1\}$.

Appendix

The Appendix consists of three parts. In Appendix A.1, we introduce some relevant definitions that are useful for the posterior analysis. The analysis of the game and the proofs of the results are presented in Appendices A.2 and A.3. In Appendix A.2, we analyze the equilibrium behavior of the wise type. This part includes the proof of Proposition 1. In Appendix A.3, we analyze the equilibrium behavior of the normal type. This part includes the proofs of all the other results in the tex.

In some cases, the analysis in the Appendix considers a more general set-up than the one described in the main body of the paper. In particular, all the results in the Appendix, except for Propositions 4 and 5, are proven for the case in which the objective function of the agent is:

$$\Pi(a, X) = f\left(\lambda_W(a, X), \lambda_{\bar{B}}(a, X)\right), \tag{3}$$

with $f(\cdot)$ denoting an increasing function in $\lambda_W(a, X)$ and $\lambda_{\bar{B}}(a, X)$. This includes the linear specification of Eq. (1) as a particular case.

Additionally, the analysis in the Appendix considers $\mu \in [0, 1]$, which includes the limit case $\mu = 0$.

A.1. Part I: Definitions

From Definition 1, the expressions of the expected gain $\Delta^{\mu}_{t,s}$ to an agent of type $t \in \{W, N\}$ for taking action R rather than L after signals L and R, respectively, when the level of transparency is $\mu \in [0, 1]$, are:

$$\begin{aligned}\Delta^{\mu}_{t,L}[\sigma_W, \sigma_N] &= \Pi^{\mu}_{t,L}(R) - \Pi^{\mu}_{t,L}(L) \\ &= (1-\mu)\Delta^{0}_{t,L}[\sigma_W, \sigma_N] + \mu\Delta^{1}_{t,L}[\sigma_W, \sigma_N] \\ &= (1-\mu)\left(f\left(\lambda_W(R,0),1\right) - f\left(\lambda_W(L,0), \lambda_{\bar{B}}(L,0)\right)\right) + \\ &\mu\left(\left(\gamma f(0,1) + (1-\gamma) f\left(\lambda_W(R,R),1\right)\right) - \left(\gamma f\left(\lambda_W(L,L), \lambda_{\bar{B}}(L,L)\right) + (1-\gamma) f\left(0, \lambda_{\bar{B}}(L,R)\right)\right)\right),\end{aligned} \tag{4}$$

$$\begin{aligned}\Delta^{\mu}_{t,R}[\sigma_W, \sigma_N] &= \Pi^{\mu}_{t,R}(R) - \Pi^{\mu}_{t,R}(L) \\ &= (1-\mu)\Delta^{0}_{t,R}[\sigma_W, \sigma_N] + \mu\Delta^{1}_{t,R}[\sigma_W, \sigma_N] \\ &= (1-\mu)\left(f\left(\lambda_W(R,0),1\right) - f\left(\lambda_W(L,0), \lambda_{\bar{B}}(L,0)\right)\right) + \\ &\mu\left(\left(\gamma f\left(\lambda_W(R,R),1\right) + (1-\gamma) f(0,1)\right) - \left(\gamma f\left(0, \lambda_{\bar{B}}(L,R)\right) + (1-\gamma) f\left(\lambda_W(L,L), \lambda_{\bar{B}}(L,L)\right)\right)\right),\end{aligned} \tag{5}$$

where we use expressions (1)–(2) and posterior $\lambda_t(a, X)$. Next, we introduce two definitions.

Definition 2. Given $\mu \in [0, 1]$, a strategy profile $\sigma^{\mu *} = (\sigma^{\mu *}_W, \sigma^{\mu *}_N)$, with $\sigma^{\mu *}_W = (\sigma^{\mu}_W(L)^*, \sigma^{\mu}_W(R)^*)$ and $\sigma^{\mu *}_N = (\sigma^{\mu}_N(L)^*, \sigma^{\mu}_N(R)^*)$ is a Perfect Bayesian equilibrium strategy profile if for each type$t \in \{W, N\}$:

1. When $s = L$, either $\Delta^{\mu}_{t,L}[\sigma^{\mu *}_W, \sigma^{\mu *}_N] = 0$ or $\Delta^{\mu}_{t,L}[\sigma^{\mu *}_W, \sigma^{\mu *}_N] > 0$ (< 0) holds. In the latter case, $\sigma^{\mu}_t(L)^* = 0$ (1).
2. When $s = R$, either $\Delta^{\mu}_{t,R}[\sigma^{\mu *}_W, \sigma^{\mu *}_N] = 0$ or $\Delta^{\mu}_{t,R}[\sigma^{\mu *}_W, \sigma^{\mu *}_N] > 0$ (< 0) holds. In the latter case, $\sigma^{\mu}_t(R)^* = 0$ (1).

Definition 2 defines an equilibrium strategy profile. To stress the fact that, in equilibrium, the strategies of the wise type and the normal type may depend on μ, we make this dependence explicit and write the superscript μ.

Definition 3. Consider $\mu = 0$. An equilibrium strategy $\bar{\sigma}^0_t(s)^*$ is robust to transparency if there exist $\bar{\mu} > 0$ and an associated equilibrium strategy $\sigma^{\bar{\mu}}_t(s)^*$ such that $\lim_{\bar{\mu}\to 0} \sigma^{\bar{\mu}}_t(s)^* = \bar{\sigma}^0_t(s)^*$.

The second definition defines a robustness criterion. This definition will be of help when analyzing the limit case $\mu = 0$, where there is a multiplicity of equilibria. To see the multiplicity, note that when $\mu = 0$, the expected gain to the agent for taking action R rather than L is the same, independent of the signal and the type, as the principal never learns the state of the world and, hence, the correct action. Mathematically, there are four variables in this case, $\sigma_W(L), \sigma_W(R), \sigma_N(L)$, and $\sigma_N(R)$, and only one equation, as $\Delta^0_{W,L}[\sigma_W, \sigma_N] = \Delta^0_{W,R}[\sigma_W, \sigma_N] = \Delta^0_{N,L}[\sigma_W, \sigma_N] = \Delta^0_{N,R}[\sigma_W, \sigma_N]$.

A.2. Part II: Analysis of the wise type

In this section, we analyze the equilibrium behavior of the wise type, for which $\gamma = 1$. For this type, Eqs. (4)–(5) simplify to:

$$\begin{aligned}\Delta^{\mu}_{W,L}[\sigma_W, \sigma_N] &= (1-\mu)\Delta^{0}_{W,L}[\sigma_W, \sigma_N] + \mu\Delta^{1}_{W,L}[\sigma_W, \sigma_N] \\ &= (1-\mu)\left(f\left(\lambda_W(R,0),1\right) - f\left(\lambda_W(L,0), \lambda_{\bar{B}}(L,0)\right)\right) + \mu\left(f(0,1) - f\left(\lambda_W(L,L), \lambda_{\bar{B}}(L,L)\right)\right),\end{aligned} \tag{6}$$

$$\begin{aligned}\Delta^{\mu}_{W,R}[\sigma_W, \sigma_N] &= (1-\mu)\Delta^{0}_{W,R}[\sigma_W, \sigma_N] + \mu\Delta^{1}_{W,R}[\sigma_W, \sigma_N] \\ &= (1-\mu)\left(f\left(\lambda_W(R,0),1\right) - f\left(\lambda_W(L,0), \lambda_{\bar{B}}(L,0)\right)\right) + \mu\left(f\left(\lambda_W(R,R),1\right) - f\left(0, \lambda_{\bar{B}}(L,R)\right)\right).\end{aligned} \tag{7}$$

Proof of Proposition 1. The results of Proposition 1 hold for the more general objective function described in (3), which is assumed in the proof.

Lemmas 1–4 prove the result. Lemma 1 introduces a preliminary technical result. Lemma 4 analyzes the behavior of the *normal* type. This lemma shows that when $\sigma_N^\mu(L)^* > 0$, the equilibrium strategy of the normal type is always informative, i.e., $\sigma_N^\mu(L)^* \neq \sigma_N^\mu(R)^*$; and that when $\sigma_N^\mu(L)^* = 0$, the equilibrium strategy of the normal type is always non-informative, i.e., $\sigma_N^\mu(L)^* = \sigma_N^\mu(R)^*$. Considering, now, the behavior of the *wise* type, Lemma 2 characterizes the equilibrium strategy of the wise type when $\sigma_N^\mu(L)^* > 0$ and obtains that, in this case, $(\sigma_W^\mu(L)^*, \sigma_W^\mu(R)^*) = (1, 0)$. Given Lemmas 4, 2 thus proves point *1.* of Proposition 1. Finally, Lemma 3 characterizes the equilibrium strategy of the wise type when $\sigma_N^\mu(L)^* = 0$. In this case, we show that there is a threshold, $\widetilde{\alpha}_B$, such that if $\alpha_B > \widetilde{\alpha}_B$, the wise type always takes action L. However, if $\alpha_B < \widetilde{\alpha}_B$, there are three equilibrium strategies, one of them being the honest strategy $(\sigma_W^\mu(L)^*, \sigma_W^\mu(R)^*) = (1, 0)$. Given Lemmas 4, 3 thus proves point *2.* of Proposition 1.

Lemma 1. *Consider $\mu > 0$. For any increasing function $f(\lambda_W(a, X), \lambda_{\bar{B}}(a, X))$, we have $\Delta_{W,L}^\mu[\sigma_W, \sigma_N] < \Delta_{N,L}^\mu[\sigma_W, \sigma_N] < \Delta_{N,R}^\mu[\sigma_W, \sigma_N] < \Delta_{W,R}^\mu[\sigma_W, \sigma_N]$.*

Proof. First, we prove that $\lambda_{\bar{B}}(L, L) \geq \lambda_{\bar{B}}(L, R)$. Note the following:

$$\lambda_B(L, L) = \frac{\alpha_B}{\alpha_B + \left(\gamma\sigma_N(L) + (1-\gamma)\sigma_N(R)\right)\alpha_N + \sigma_W(L)\alpha_W},$$

$$\lambda_B(L, R) = \frac{\alpha_B}{\alpha_B + \left((1-\gamma)\sigma_N(L) + \gamma\sigma_N(R)\right)\alpha_N + \sigma_W(R)\alpha_W},$$

with $\lambda_B(L, L) \leq \lambda_B(L, R) \iff (1-\gamma)\sigma_N(L) + \gamma\sigma_N(R) \leq \gamma\sigma_N(L) + (1-\gamma)\sigma_N(R) \iff (1-2\gamma)\sigma_N(L) \leq (1-2\gamma)\sigma_N(R) \iff \sigma_N(L) \geq \sigma_N(R)$. This is always the case in a non-perverse equilibrium. For the same reason, $\sigma_W(L) \geq \sigma_W(R)$. Hence, $\lambda_B(L, L) \leq \lambda_B(L, R)$, and consequently, $\lambda_{\bar{B}}(L, L) \geq \lambda_{\bar{B}}(L, R)$.

Second, note that $\lambda_{\bar{B}}(L, L) \geq \lambda_{\bar{B}}(L, R)$ implies $f\left(\lambda_W(L, L), \lambda_{\bar{B}}(L, L)\right) > f\left(0, \lambda_{\bar{B}}(L, R)\right)$. Finally, from Eqs. (4)–(5) and (6)–(7), it is straightforward to show the following:

$$f\left(\lambda_W(L, L), \lambda_{\bar{B}}(L, L)\right) > f\left(0, \lambda_{\bar{B}}(L, R)\right) \Longrightarrow \Delta_{W,L}^\mu[\sigma_W, \sigma_N] < \Delta_{N,L}^\mu[\sigma_W, \sigma_N],$$

$$f\left(\lambda_W(L, L), \lambda_{\bar{B}}(L, L)\right) > f\left(0, \lambda_{\bar{B}}(L, R)\right) \Longrightarrow \Delta_{N,L}^\mu[\sigma_W, \sigma_N] < \Delta_{N,R}^\mu[\sigma_W, \sigma_N],$$

$$f\left(\lambda_W(L, L), \lambda_{\bar{B}}(L, L)\right) > f\left(0, \lambda_{\bar{B}}(L, R)\right) \Longrightarrow \Delta_{N,R}^\mu[\sigma_W, \sigma_N] < \Delta_{W,R}^\mu[\sigma_W, \sigma_N]. \quad \blacklozenge$$

Lemma 2. *Consider $\mu > 0$ and $\sigma_N^\mu(L)^* > 0$. For any increasing function $f(\lambda_W(a, X), \lambda_{\bar{B}}(a, X))$, the unique equilibrium strategy of the wise type is $(\sigma_W^\mu(L)^*, \sigma_W^\mu(R)^*) = (1, 0)$.*

Proof. First, we prove that $\sigma_W^\mu(R)^* = 0$. We prove it by contradiction, so let us assume $\sigma_W^\mu(R)^* > 0$. In this case, $\Delta_{W,R}^\mu[\sigma_W^*, \sigma_N^*] \leq 0$; hence, by Lemma 1, $\Delta_{W,L}^\mu[\sigma_W^*, \sigma_N^*] < \Delta_{N,L}^\mu[\sigma_W^*, \sigma_N^*] < \Delta_{N,R}^\mu[\sigma_W^*, \sigma_N^*] < \Delta_{W,R}^\mu[\sigma_W^*, \sigma_N^*] \leq 0$, and therefore, $\sigma_W^\mu(L)^* = 1$, $\sigma_N^\mu(L)^* = 1$ and $\sigma_N^\mu(R)^* = 1$. This implies that if the wise type takes action R, the principal's equilibrium beliefs assign probability one to the agent being wise and zero probability to her being biased. Thus, the payoff to the agent for taking action R is always higher than the payoff from taking action L. Hence, $\sigma_W^\mu(R)^*$ cannot be greater than zero in equilibrium, which is a contradiction. Consequently, in equilibrium, $\sigma_W^\mu(R)^* = 0$ always.

Second, we prove that $\sigma_W^\mu(L)^* = 1$. Note that if $\sigma_N^\mu(L)^* > 0$, then $\Delta_{N,L}^\mu[\sigma_W^*, \sigma_N^*] \leq 0$. Now, by Lemma 1, $\Delta_{W,L}^\mu[\sigma_W^*, \sigma_N^*] < \Delta_{N,L}^\mu[\sigma_W^*, \sigma_N^*] \leq 0$. Then, in equilibrium, $\sigma_W^\mu(L)^* = 1$ always. $\blacklozenge$

Lemma 3. *Consider $\mu > 0$ and $\sigma_N^\mu(L)^* = 0$. For any increasing function $f(\lambda_W(a, X), \lambda_{\bar{B}}(a, X))$, in equilibrium $\sigma_N^\mu(R)^* = \sigma_W^\mu(R)^* = 0$. In the case in which $s = L$, there exists threshold $\widetilde{\alpha}_B$ such that if $\alpha_B > \widetilde{\alpha}_B$, then $\sigma_W^\mu(L)^* = 0$; and if $\alpha_B < \widetilde{\alpha}_B$, then in equilibrium either $\sigma_W^\mu(L)^* = 0$, $\sigma_W^\mu(L)^* = 1$ or $\sigma_W^\mu(L)^* \in (0, 1)$. The described equilibrium strategies are the unique ones.*

Proof. First, note that if $\sigma_N^\mu(L)^* = 0$, then $\Delta_{N,L}^\mu[\sigma_W^*, \sigma_N^*] \geq 0$. Hence, by Lemma 1, $0 \leq \Delta_{N,L}^\mu[\sigma_W^*, \sigma_N^*] < \Delta_{N,R}^\mu[\sigma_W^*, \sigma_N^*] < \Delta_{W,R}^\mu[\sigma_W^*, \sigma_N^*]$. It implies $\sigma_N^\mu(R)^* = 0$ and $\sigma_W^\mu(R)^* = 0$.

To obtain the equilibrium value $\sigma_W^\mu(L)^*$, we analyze Eq. (6). First, note that when $\sigma_N^\mu(L)^* = 0$, $\sigma_N^\mu(R)^* = 0$ and $\sigma_W^\mu(R)^* = 0$, beliefs are as follows:

$$\lambda_W(R, 0) = \frac{\left(2 - \sigma_W(L)\right)\alpha_W}{\left(2 - \sigma_W(L)\right)\alpha_W + 2\left(1 - \alpha_W - \alpha_B\right)},$$

$$\lambda_W(L, 0) = \lambda_{\bar{B}}(L, 0) = \frac{\sigma_W(L)\alpha_W}{\sigma_W(L)\alpha_W + 2\alpha_B},$$

$$\lambda_W(L, L) = \lambda_{\bar{B}}(L, L) = \frac{\sigma_W(L)\alpha_W}{\sigma_W(L)\alpha_W + \alpha_B},$$

with $\frac{\partial \lambda_W(R,0)}{\partial \sigma_W(L)} < 0$, $\frac{\partial \lambda_W(L,0)}{\partial \sigma_W(L)} > 0$ and $\frac{\partial \lambda_W(L,L)}{\partial \sigma_W(L)} > 0$. Consequently, $\frac{\partial \Delta^{\mu}_{W,L}[\sigma_W,\sigma_N]}{\partial \sigma_W(L)} < 0$.

Additionally, $\Delta^{\mu}_{W,L}\Big|_{\sigma_W(L)=0} = (1-\mu)\left(f\left(\frac{\alpha_W}{\alpha_W+(1-\alpha_W-\alpha_B)},1\right) - f(0,0)\right) + \mu(f(0,1)-f(0,0)) > 0$.

Then, the sign of $\Delta^{\mu}_{W,L}\Big|_{\sigma_W(L)=1}$ determines whether there is either one equilibrium, $\sigma^{\mu}_W(L)^* = 0$, or three equilibria: $\sigma^{\mu}_W(L)^* = 0$, $\sigma^{\mu}_W(L)^* = 1$ and $\sigma^{\mu}_W(L)^* \in (0,1)$. Note that if $\Delta^{\mu}_{W,L}\Big|_{\sigma_W(L)=1} \leq 0$, there are three equilibria, and if $\Delta^{\mu}_{W,L}\Big|_{\sigma_W(L)=1} > 0$, there is only one equilibrium.

Note that

$$\Delta^{\mu}_{W,L}\Big|_{\sigma_W(L)=1} = (1-\mu)\left(f\left(\frac{\alpha_W}{\alpha_W+2\left(1-\alpha_W-\alpha_B\right)},1\right) - f\left(\frac{\alpha_W}{\alpha_W+2\alpha_B},\frac{\alpha_W}{\alpha_W+2\alpha_B}\right)\right) + \mu\left(f(0,1) - f\left(\frac{\alpha_W}{\alpha_W+\alpha_B},\frac{\alpha_W}{\alpha_W+\alpha_B}\right)\right). \tag{8}$$

This expression is increasing in α_B. Additionally, when α_B goes to zero, we have $\Delta^{\mu}_{W,L}\Big|_{\sigma_W(L)=1} = (1-\mu)(f(\frac{\alpha_W}{2-\alpha_W},1) - f(1,1)) + \mu(f(0,1)-f(1,1))$, which is negative since $f(\cdot)$ is increasing in both arguments. See expression (3). Similarly, when α_B goes to one, we have $\Delta^{\mu}_{W,L}\Big|_{\sigma_W(L)=1} = (1-\mu)(f(0,1)-f(0,0)) + \mu(f(0,1)-f(0,0))$, which is positive because of the same reason. Consequently, there exists a threshold for α_B, denoted by $\widetilde{\alpha}_B$, such that above the threshold, there is a unique equilibrium, and below the threshold, there are three equilibria. ♦

Lemma 4. *In equilibrium,*

1. *If $\sigma^{\mu}_N(L)^* = 0$, then the unique equilibrium strategy of the normal type is non-informative, i.e., $\sigma^{\mu}_N(L)^* = \sigma^{\mu}_N(R)^*$.*
2. *If $\sigma^{\mu}_N(L)^* > 0$, then the equilibrium strategy of the normal type is informative, i.e., $\sigma^{\mu}_N(L)^* \neq \sigma^{\mu}_N(R)^*$.*

Proof. Case 1. From Lemma 3, we know that if $\sigma^{\mu}_N(L)^* = 0$, then in equilibrium, $\sigma^{\mu}_N(R)^* = 0$ always. The strategy of the normal type is thus non-informative.

Case 2. From Lemma 2, we know that if $\sigma^{\mu}_N(L)^* > 0$, then in equilibrium, $(\sigma^{\mu}_W(L)^*, \sigma^{\mu}_W(R)^*) = (1,0)$ always. Additionally, Proposition 2 in Appendix A.3 below shows that if $(\sigma^{\mu}_W(L)^*, \sigma^{\mu}_W(R)^*) = (1,0)$, then in equilibrium, $\sigma^{\mu}_N(R)^* = 0$ always. Combining these results, we have that the equilibrium strategy of the normal type is informative if and only if $\sigma^{\mu}_N(L)^* > 0$.

Hence, in our model, there is an equivalence between stating $\sigma^{\mu}_N(L)^* > 0$ and saying that the normal type uses an informative strategy, and between stating $\sigma^{\mu}_N(L)^* = 0$ and saying that the normal type uses a non-informative strategy. ♦

This completes the proof of Proposition 1. □

Remark on Proposition 1. When $\mu = 0$, then agent's expected gain for taking action R rather than L is always the same, independently of the signal and the type, since the principal never learns the state. Because there are four variables in this case, $\sigma_W(L)$, $\sigma_W(R)$, $\sigma_N(L)$, and $\sigma_N(R)$, and only one equation, $\Delta^0_{W,L}[\sigma_W,\sigma_N] = \Delta^0_{N,L}[\sigma_W,\sigma_N] = \Delta^0_{N,R}[\sigma_W,\sigma_N] = \Delta^0_{W,R}[\sigma_W,\sigma_N]$, there are multiple equilibria. Among the multiplicity, the following strategy profiles of the wise type $(\sigma^{\mu}_W(L)^*, \sigma^{\mu}_W(R)^*) = (0,0)$, $(\sigma^{\mu}_W(L)^*, \sigma^{\mu}_W(R)^*) = (1,0)$, and $(\sigma^{\mu}_W(L)^*, \sigma^{\mu}_W(R)^*) = (x_1,0)$, with $x_1 \in (0,1)$, are equilibrium strategies when $\mu = 0$. Additionally, from Proposition 1, they are the unique equilibrium strategies when $\mu > 0$. Hence, by Definition 3, they are the only equilibrium strategies of the wise type that are robust to transparency when $\mu = 0$.

A.3. Part III: Analysis of the normal type

First, note that Proposition 1 above shows that if the normal type uses an informative strategy, then the unique equilibrium strategy of the wise type is $(\sigma^{\mu}_W(L)^*, \sigma^{\mu}_W(R)^*) = (1,0)$. Hence, if the analysis that follows shows that under this premise the equilibrium strategy of the normal type is informative, then we have an informative equilibrium. However, if the best response of the normal type is to use a non-informative strategy, then to know whether there is a (non-informative) equilibrium in which the normal type uses a non-informative strategy, we need to analyze the following: (i) the best response of the wise type to a normal type using a non-informative strategy and (ii) the best response of the normal type to the previous optimal response of the wise type. Note that Proposition 1 completely characterizes the wise type's optimal behavior in point (i). It shows that if $\alpha_B > \widetilde{\alpha}_B$, the unique equilibrium strategy of the wise type is $(\sigma^{\mu}_W(L)^*, \sigma^{\mu}_W(R)^*) = (0,0)$. However, if $\alpha_B < \widetilde{\alpha}_B$, the wise type has three equilibrium strategies that can be subsumed into the following two: $(\sigma^{\mu}_W(L)^*, \sigma^{\mu}_W(R)^*) = (1,0)$ and $(\sigma^{\mu}_W(L)^*, \sigma^{\mu}_W(R)^*) = (x_1,0)$, with $x_1 < 1$. The next two results, Proposition 8 and Corollary 1, characterize the equilibrium strategy of the normal type when the wise type does not always follow her signal, i.e., when $(\sigma^{\mu}_W(L)^*, \sigma^{\mu}_W(R)^*) \neq (1,0)$. Hence, these results answer point (ii).

Proposition 8. *Consider $\sigma^{\mu}_W(L)^* < 1$ and $\mu \in [0,1]$. For any increasing function $f(\lambda_W(a,X), \lambda_{\bar{B}}(a,X))$, in equilibrium, $\sigma^{\mu}_W(R)^* = 0$ and $(\sigma^{\mu}_N(L)^*, \sigma^{\mu}_N(R)^*) = (0,0)$. Furthermore, if $\mu > 0$, the equilibrium strategy of the normal type is unique.*

Proof. The proof restricts attention to the case $\mu > 0$. Note that if $\mu = 0$, there is a multiplicity of equilibria, one of which prescribes $\sigma^{\mu}_W(R)^* = 0$ and $(\sigma^{\mu}_N(L)^*, \sigma^{\mu}_N(R)^*) = (0,0)$.

Now, if $\sigma_W^\mu(L)^* < 1$, then $\Delta_{W,L}^\mu[\sigma_W^*, \sigma_N^*] \geq 0$. By Lemma 1, $0 < \Delta_{N,L}^\mu[\sigma_W^*, \sigma_N^*] < \Delta_{N,R}^\mu[\sigma_W^*, \sigma_N^*] < \Delta_{W,R}^\mu[\sigma_W^*, \sigma_N^*]$, which implies $\sigma_W^\mu(R)^* = 0$, $\sigma_N^\mu(L)^* = 0$, and $\sigma_N^\mu(R)^* = 0$. □

Proof of Corollary 1. Note that if $(\sigma_W^\mu(L)^*, \sigma_W^\mu(R)^*) \neq (1,0)$, then necessarily either $\sigma_W^\mu(L)^* < 1$ or $\sigma_W^\mu(R)^* > 0$. By Proposition 1, there is no equilibrium in which $\sigma_W^\mu(R)^* > 0$; hence, $\sigma_W^\mu(L)^* < 1$ must hold. Now, by Proposition 8, if $\sigma_W^\mu(L)^* < 1$, then $(\sigma_N^\mu(L)^*, \sigma_N^\mu(R)^*) = (0,0)$. □

Proof of Proposition 2. The result of Proposition 2 holds for the more general objective function described in (3), which is assumed in the proof.

We start noting that, by Corollary 1, if the wise type does not use the honest strategy, then $\sigma_N^\mu(R)^* = 0$. Hence, we next consider that the wise type uses the honest strategy, i.e., $(\sigma_W^\mu(L)^*, \sigma_W^\mu(R)^*) = (1,0)$.

In this case, we first prove that, in equilibrium, the normal type cannot disregard both signals L and R with positive probability. We prove this by contradiction. Hence, suppose that $\sigma_N^\mu(L)^* < 1$ and $\sigma_N^\mu(R)^* > 0$. Since $\sigma_N^\mu(R)^* > 0$, then $\Delta_{N,R}^\mu \leq 0$. Additionally, since $\sigma_N^\mu(L)^* < 1$, then $\Delta_{N,L}^\mu \geq 0$. This contradicts Lemma 1, which states $\Delta_{N,L}^\mu < \Delta_{N,R}^\mu$. Then, the equilibrium strategy profile of the normal type is either $(\sigma_N^\mu(L)^*, \sigma_N^\mu(R)^*) = (y_1, 0)$ with $y_1 < 1$, or $(\sigma_N^\mu(L)^*, \sigma_N^\mu(R)^*) = (1, y_2)$ with $y_2 > 0$.

Next, we show that $\sigma_N^\mu(R)^* > 0$ is not possible. Again, we prove it by contradiction. Suppose that $\sigma_N^\mu(R)^* > 0$. Then, $\sigma_N^\mu(L)^* = 1$. Now, we use the results (9) and (10), which we prove below:

$$\lambda_W(R,R) > \lambda_W(L,L) \iff (1 - \sigma_N(L) - \sigma_N(R))\alpha_N < \alpha_B, \tag{9}$$

$$\lambda_W(R,0) > \lambda_W(L,0) \iff (1 - \sigma_N(L) - \sigma_N(R))\alpha_N < \alpha_B. \tag{10}$$

In the case that $\sigma_N(R) > 0$ and $\sigma_N(L) = 1$, then $\lambda_W(R,R) > \lambda_W(L,L)$ and $\lambda_W(R,0) > \lambda_W(L,0)$. Substituting in expression (5), we obtain $\Delta_{N,R}^\mu > 0$. However, $\Delta_{N,R}^\mu > 0$ implies $\sigma_N^\mu(R)^* = 0$, which contradicts $\sigma_N^\mu(R)^* > 0$. As a result, in equilibrium, $\sigma_N^\mu(R)^* = 0$.

To complete the proof, we prove conditions (9) and (10). Given the equilibrium strategy $(\sigma_W^\mu(L)^*, \sigma_W^\mu(R)^*) = (1,0)$, Bayes' rule determines the following:

$$\lambda_W(R,R) = \frac{\alpha_W}{\alpha_W + (\gamma(1 - \sigma_N(R)) + (1-\gamma)(1 - \sigma_N(L)))\alpha_N} \qquad \lambda_W(R,0) = \frac{\alpha_W}{\alpha_W + (2 - \sigma_N(R) - \sigma_N(L))\alpha_N}$$

$$\lambda_W(L,L) = \frac{\alpha_W}{\alpha_B + \alpha_W + ((1-\gamma)\sigma_N(R) + \gamma\sigma_N(L))\alpha_N} \qquad \lambda_W(L,0) = \frac{\alpha_W}{2\alpha_B + \alpha_W + (\sigma_N(R) + \sigma_N(L))\alpha_N}$$

First, we observe that $\lambda_W(R,R) > \lambda_W(L,L) \iff \alpha_B + \alpha_W + \big((1-\gamma)\sigma_N(R) + \gamma\sigma_N(L)\big)\alpha_N > \alpha_W + \big(\gamma(1 - \sigma_N(R)) + (1-\gamma)(1 - \sigma_N(L))\big)\alpha_N \iff (1 - \sigma_N(L) - \sigma_N(R))\alpha_N < \alpha_B$.

Second, we observe that $\lambda_W(R,0) > \lambda_W(L,0) \iff 2\alpha_B + \alpha_W + (\sigma_N(L) + \sigma_N(R))\alpha_N > \alpha_W + (2 - \sigma_N(L) - \sigma_N(R))\alpha_N \iff (1 - \sigma_N(L) - \sigma_N(R))\alpha_N < \alpha_B$. □

Remark on Proposition 2. When $\mu = 0$, there are multiple equilibria. According to Definition 3, in all equilibria that are robust to transparency, $\sigma_N^\mu(R)^* = 0$.

Proof of Proposition 3. The results of Proposition 3 hold for the more general objective function described in (3), which is assumed in the proof.

First, we prove point *1.* of the proposition. To show that when $\alpha_B > \alpha_B^{max}$, $\Delta_{N,L}^\mu > 0$ for any $\mu \in [0,1]$, we show that there exist two thresholds $\alpha_B^0 < 1$ and $\alpha_B^1 < 1$, such that $\Delta_{N,L}^0 > 0$ for all $\alpha_B > \alpha_B^0$ and $\Delta_{N,L}^1 > 0$ for all $\alpha_B > \alpha_B^1$.

Let us start with α_B^1. Note that $\Delta_{N,L}^1 = \Pi_{N,L}^1(R) - \Pi_{N,L}^1(L)$, with $\Pi_{N,L}^1(L) = \gamma f\left(\lambda_W(L,L), \lambda_{\bar{B}}(L,L)\right) + (1-\gamma) f\left(0, \lambda_{\bar{B}}(L,R)\right)$ and $\Pi_{N,L}^1(R) = \gamma f(0,1) + (1-\gamma) f\left(\lambda_W(R,R), 1\right) > \gamma f(0,1) + (1-\gamma) f(0,1) = f(0,1) > 0$; the latter implying $\Pi_{N,L}^1(R) > 0$ always. It can be shown that $\Pi_{N,L}^1(L)$ is decreasing in α_B (either with α_W or α_N constant), with $lim_{\alpha_B \to 1} \Pi_{N,L}^1(L) \to 0$ (as beliefs go to zero when α_B goes to one). Hence, there always exists $\alpha_B^1 < 1$ such that for any $\alpha_B > \alpha_B^1$, $\Pi_{N,L}^1(R) > \Pi_{N,L}^1(L)$, and consequently, $\Delta_{N,L}^1 > 0$.

Now, to prove the existence of α_B^0, first note that $\Delta_{N,L}^0 = f\left(\lambda_W(R,0), 1\right) - f\left(\lambda_W(L,0), \lambda_{\bar{B}}(L,0)\right)$. If the wise type plays the honest strategy, i.e. $(\sigma_W^\mu(L)^*, \sigma_W^\mu(R)^*) = (1,0)$, a sufficient condition for $\Delta_{N,L}^0 > 0$ is $\lambda_W(R,0) > \lambda_W(L,0)$, which is satisfied if and only if $(1 - \sigma_N(L) - \sigma_N(R))\alpha_N < \alpha_B$ (see condition (10)). Now, since $(1 - \sigma_N(L) - \sigma_N(R))\alpha_N < 1$, there always exists $\alpha_B^0 < 1$ for which $(1 - \sigma_N(L) - \sigma_N(R))\alpha_N < \alpha_B^0$. Consequently, for any $\alpha_B > \alpha_B^0$, $\Delta_{N,L}^0 > 0$.

Therefore, if the wise type plays the honest strategy, then for all $\alpha_B > \max\{\alpha_B^0, \alpha_B^1\}$, $\Delta_{N,L}^\mu = (1-\mu)\Delta_{N,L}^0 + \mu\Delta_{N,L}^1 > 0$ for any $\mu \in [0,1]$, which implies that $\sigma_N^\mu(L)^* = 0$. Thus, in this case, the unique equilibrium strategy of the normal type is $(\sigma_N^\mu(L)^*, \sigma_N^\mu(R)^*) = (0,0)$. Note also that if the wise type does not play the honest strategy, i.e. $(\sigma_W^\mu(L)^*, \sigma_W^\mu(R)^*) \neq (1,0)$, then by Corollary 1, the unique equilibrium strategy of the normal type is $(\sigma_N^\mu(L)^*, \sigma_N^\mu(R)^*) = (0,0)$. Consequently, in this case, if $\alpha_B > \max\{\alpha_B^0, \alpha_B^1\}$, then for any strategy of the wise type, the unique equilibrium strategy of the normal type is $(\sigma_N^\mu(L)^*, \sigma_N^\mu(R)^*) = (0,0)$.

To complete the proof of point *1.* note that, by Lemma 3, if the strategy of the normal type is $(\sigma_N^\mu(L)^*, \sigma_N^\mu(R)^*) = (0,0)$, we know that there exists threshold $\widetilde{\alpha}_B$ for the wise type such that, for all $\alpha_B > \widetilde{\alpha}_B$, the wise type always takes action R. Let $\alpha_B^{max} = \max\{\alpha_B^0, \alpha_B^1, \widetilde{\alpha}_B\}$, and note that $\alpha_B^{max} = \widetilde{\alpha}_B$ (otherwise, we could have a situation in which the normal type uses an informative

strategy and the wise type uses a non-informative strategy, which contradicts point *1.* of Proposition 1). Therefore, for all $\alpha_B > \alpha_B^{max}$, there is a unique equilibrium in which both the normal type and the wise type always take action R.

Second, we prove point *2.* of the proposition. First note that by Corollary 1, in any informative equilibrium the wise type always uses the honest strategy, i.e. $(\sigma_W^\mu(L)^*, \sigma_W^\mu(R)^*) = (1,0)$. Thus, we consider $(\sigma_W^\mu(L)^*, \sigma_W^\mu(R)^*) = (1,0)$ and show that for all $\mu \in [0,1]$ and $\alpha_B < \alpha_B^{min}$, $\Delta_{N,L}^\mu < 0$. First, note that condition $\Delta_{N,L}^\mu = (1-\mu)\Delta_{N,L}^0 + \mu\Delta_{N,L}^1 < 0$ is equivalent to $\Delta_{N,L}^0 < \frac{\mu}{1-\mu}(-\Delta_{N,L}^1)$. We use two lemmas to prove the result: Lemmas 5 and 6, which are stated and proven below. Lemma 5 shows that for $\varepsilon_1 > 0$, there always exists $\alpha_{B(\varepsilon_1)} > 0$ such that for any $\alpha_B < \alpha_{B(\varepsilon_1)}$, in equilibrium $\Delta_{N,L}^0 < \varepsilon_1$. Lemma 6 shows that there exist $\varepsilon_2 > 0$ and $\alpha_{B(\varepsilon_2)} > 0$ such that for any $\alpha_B < \alpha_{B(\varepsilon_2)}$, $\Delta_{N,L}^1 < -\varepsilon_2$. Therefore, if we take $\varepsilon_1 < \frac{\mu}{1-\mu}\varepsilon_2$ and $\alpha_B^{min} = \min\{\alpha_{B(\varepsilon_1)}, \alpha_{B(\varepsilon_2)}, \widetilde{\alpha}_B\}$, then for any $\alpha_B < \alpha_B^{min}$, we have that $\Delta_{N,L}^0 < \varepsilon_1 < \frac{\mu}{1-\mu}\varepsilon_2 < \frac{\mu}{1-\mu}(-\Delta_{N,L}^1) \iff \Delta_{N,L}^\mu = (1-\mu)\Delta_{N,L}^0 + \mu\Delta_{N,L}^1 < 0$, and consequently, $\sigma_N^\mu(L)^* = 1$.

Lemma 5. *For any $\varepsilon_1 > 0$, there always exists $\alpha_{B(\varepsilon_1)} > 0$ such that $\forall \alpha_B < \alpha_{B(\varepsilon_1)}$, in equilibrium, $\Delta_{N,L}^0 < \varepsilon_1$.*

Proof. Note that $\Delta_{N,L}^0 = f\left(\lambda_W(R,0),1\right) - f\left(\lambda_W(L,0), \lambda_{\bar{B}}(L,0)\right)$. From the beliefs in Table 2, we have $\frac{\partial \lambda_{\bar{B}}(L,0)}{\partial \alpha_B} < 0$ and $lim_{\alpha_B \to 0}\lambda_{\bar{B}}(L,0) \to 1$, so $\Delta_{N,L}^0 = f\left(\lambda_W(R,0),1\right) - f\left(\lambda_W(L,0),1\right)$ for α_B sufficiently small. Additionally, from the beliefs in Table 1, we have $lim_{\alpha_B \to 0}\lambda_W(R,0) \leq lim_{\alpha_B \to 0}\lambda_W(L,0) \iff \alpha_W + \sigma_N(L)\alpha_N \leq \alpha_W + (2-\sigma_N(L))\alpha_N \iff \sigma_N(L) \leq 1$; so it always holds. Hence, $lim_{\alpha_B \to 0}\Delta_{N,L}^0 \leq 0$. Then, we can assert there exists $\alpha_{B(\varepsilon_1)} > 0$, such that for any $\alpha_B < \alpha_{B(\varepsilon_1)}$, $\Delta_{N,L}^0 < \varepsilon_1$. ♦

Lemma 6. *There exists $\varepsilon_2 > 0$ and $\alpha_{B(\varepsilon_2)} > 0$, such that $\forall \alpha_B < \alpha_{B(\varepsilon_2)}$, $\Delta_{N,L}^1 < -\varepsilon_2$.*

Proof. Note that $\Delta_{N,L}^1 = \gamma f(0,1) + (1-\gamma) f\left(\lambda_W(R,R),1\right) - \left(\gamma f\left(\lambda_W(L,L), \lambda_{\bar{B}}(L,L)\right) + (1-\gamma) f\left(0, \lambda_{\bar{B}}(L,R)\right)\right)$. By Proposition 2, $\sigma_N^\mu(R)^* = 0$. Assuming $\sigma_N^\mu(R)^* = 0$ and taking limits on the beliefs in Table 1, we obtain $lim_{\alpha_B \to 0}\lambda_W(R,R) \geq \alpha_W > 0$ and $lim_{\alpha_B \to 0}\lambda_W(L,L) \geq \alpha_W > 0$. Note also that $lim_{\alpha_B \to 0}\lambda_W(R,R) > lim_{\alpha_B \to 0}\lambda_W(L,L) \iff \gamma + (1-\gamma)(1-\sigma_N(L)) > \gamma\sigma_N(L)$. Therefore, if $\sigma_N(L) < 1$, then $lim_{\alpha_B \to 0}\lambda_W(R,R) > lim_{\alpha_B \to 0}\lambda_W(L,L)$, and if $\sigma_N(L) = 1$, then $lim_{\alpha_B \to 0}\lambda_W(R,R) = lim_{\alpha_B \to 0}\lambda_W(L,L)$. Let us denote $k = lim_{\alpha_B \to 0}\lambda_W(R,R)$ and $K = lim_{\alpha_B \to 0}\lambda_W(L,L)$.

Now, taking limits on the beliefs in Table 2, we obtain $lim_{\alpha_B \to 0}\lambda_{\bar{B}}(L,L) \to 1$ and $lim_{\alpha_B \to 0}\lambda_{\bar{B}}(L,R) \to 1$.

Using these results, we have $lim_{\alpha_B \to 0}\Delta_{N,L}^1 = \gamma f(0,1) + (1-\gamma) f(k,1) - (\gamma f(K,1) + (1-\gamma) f(0,1))$. Let $z = lim_{\alpha_B \to 0}\Delta_{N,L}^1$. First, note that $\gamma f(0,1) + (1-\gamma) f(k,1) < \gamma f(K,1) + (1-\gamma) f(0,1)$, as $\gamma > \frac{1}{2}$ and $f(K,1) \geq f(k,1)$. Thus, $z < 0$. Now, let $\varepsilon_2 > 0$ be any number in the interval $(0,|z|)$. Then, $lim_{\alpha_B \to 0}\Delta_{N,L}^1 < -\varepsilon_2$. Thus, there always exists $\alpha_{B(\varepsilon_2)} > 0$, such that if $\alpha_B < \alpha_{B(\varepsilon_2)}$, then $\Delta_{N,L}^1 < -\varepsilon_2$. ♦

This completes the proof of Proposition 3. □

Remark 1 on Proposition 3. When $\alpha_B > \alpha_B^{max}$ and $\mu = 0$, there are multiple equilibria. According to Definition 3, in all the equilibria that are robust to transparency, we have $\sigma_N^\mu(L)^* = 0$.

Remark 2 on Proposition 3. When $\mu = 0$, in equilibrium, we always have $\sigma_N^\mu(L)^* < 1$. Briefly, when $\mu \to 0$, $\alpha_B^{min} \to 0$. Therefore, there is no α_B that is low enough to sustain an equilibrium in which $\sigma_N^\mu(L)^* = 1$.

Proof of Proposition 4. In the proof, we use two simplifications, which are without loss of generality:

- We consider $\eta = \beta = 1$ and prove the result for this case. Note that $\Delta_{N,L}^\mu \geq 0$ when $\eta = \beta = 1$ if and only if $\Delta_{N,L}^\mu \geq 0$ when $\eta = \beta$. See expression (4) with $f\left(\lambda_W(a,X), \lambda_{\bar{B}}(a,X)\right) = \eta\lambda_W(a,X) + \beta\lambda_{\bar{B}}(a,X)$. Hence, the analysis that follows proves the result for any $\eta = \beta > 0$.

- We use the result in Proposition 2 and Definition 3 and, therefore, consider $\sigma_N^\mu(R)^* = 0\ \forall \mu \in [0,1]$.

We start noting that, by Corollary 1, if the wise type does not use the honest strategy, then $(\sigma_N^\mu(L)^*, \sigma_N^\mu(R)^*) = (0,0)$ for all μ. Hence, if we want to analyze whether an informative equilibrium exists, we have to consider that the wise type uses the honest strategy, i.e., $(\sigma_W^\mu(L)^*, \sigma_W^\mu(R)^*) = (1,0)$. This is what we do in points *1.* and *2.* of Proposition 4. It is also what we assume in order to prove point *3.* of the Proposition. The difference is that, in this case, we obtain that the equilibrium strategy of the normal type is $(\sigma_N^\mu(L)^*, \sigma_N^\mu(R)^*) = (0,0)$, which is non-informative. To complete the proof of this point, we show that $\alpha_B'' < \widetilde{\alpha}_B$ (see Lemma 10 below). Then, by Proposition 1, in equilibrium either $(\sigma_W^\mu(L)^*, \sigma_W^\mu(R)^*) = (0,0)$, $(\sigma_W^\mu(L)^*, \sigma_W^\mu(R)^*) = (1,0)$ or $(\sigma_W^\mu(L)^*, \sigma_W^\mu(R)^*) = (x_1,0)$, with $x_1 \in (0,1)$.

Next, we move on to prove Proposition 4, which requires a thorough analysis of expression $\Delta_{N,L}^\mu = (1-\mu)\Delta_{N,L}^0 + \mu\Delta_{N,L}^1$. This analysis is structured in three lemmas: Lemmas 7–9.

Lemma 8 analyzes the behavior of $\Delta_{N,L}^\mu$ when $\mu = 1$. It shows that $\Delta_{N,L}^1 > 0$ for any $\alpha_B \in (\alpha_B', \alpha_B'')$ and $\gamma \in (\frac{1}{2}, \gamma')$, with $\alpha_B' = \frac{\alpha_W\left(2+\alpha_W+\alpha_W^2-2\sqrt{2}\sqrt{\alpha_W(\alpha_W+1)}\right)}{(\alpha_W+2)^2}$, $\alpha_B'' = \frac{2+\alpha_W-\sqrt{9\alpha_W^2-4\alpha_W+4}}{4}$ and $\gamma' = \frac{\left(\alpha_B^2-\alpha_W^2\right)}{2\sqrt{2\alpha_B\alpha_W}-(2\alpha_B+\alpha_W)(1-\alpha_B+\alpha_W)}$. Lemma 7 focuses on the analysis of $\Delta_{N,L}^\mu$ when $\mu = 0$. It shows that when $\alpha_B < \alpha_B''$, hence when $\alpha_B \in \left(\alpha_B', \alpha_B''\right)$, the following occurs: *(1)*: $\frac{\partial \Delta_{N,L}^0}{\partial \sigma_N(L)} > 0$, *(2)*: $\Delta_{N,L}^0\Big|_{\sigma_N(L)=0} < 0$, *(3)*: $\Delta_{N,L}^0\Big|_{\sigma_N(L)=1} > 0$, and *(4)*: $\Delta_{N,L}^0\Big|_{\sigma_N^0(L)^*} = 0$, with $\sigma_N^0(L)^* = 1 - \frac{2\alpha_B\alpha_W+\alpha_B\alpha_N}{(\alpha_W-\alpha_B)\alpha_N}$. Consequently, $\Delta_{N,L}^0 < 0$ if $\sigma_N(L) < \sigma_N^0(L)^*$, and $\Delta_{N,L}^0 > 0$ if $\sigma_N(L) > \sigma_N^0(L)^*$. Finally, Lemma 9 focuses on $\sigma_N^\mu(L)^{Sup*}$, which denotes the highest equilibrium probability with which the normal type takes action L after signal L. Note that in the most informative equilibrium, $\sigma_N^\mu(L)^* = \sigma_N^\mu(L)^{Sup*}$. We refer to $\sigma_N^\mu(L)^{Sup*}$ as the supremum. This lemma shows that $\sigma_N^\mu(L)^{Sup*}$ decreases in μ.

The results in these lemmas allow us to state that if $\alpha_B \in (\alpha'_B, \alpha''_B)$ and $\gamma \in \left(\frac{1}{2}, \gamma'\right)$:

(i) There always exist $\mu'' < 1$ such that if $\mu > \mu''$, $\Delta^{\mu}_{N,L} = (1-\mu)\Delta^0_{N,L} + \mu\Delta^1_{N,L} > 0$ and if $\mu < \mu''$, $\Delta^{\mu}_{N,L}$ has, at least, one root. This is so because $\Delta^1_{N,L} > 0$ and $\Delta^0_{N,L}$ is increasing in $\sigma_N(L)$, with $\Delta^0_{N,L}\Big|_{\sigma_N(L)=0} < 0$ and $\Delta^0_{N,L}\Big|_{\sigma_N(L)=1} > 0$. This implies that, if $\mu > \mu''$, in the unique equilibrium strategy of the normal type, $\sigma^{\mu}_N(L)^* = 0$. This result describes the equilibrium behavior of the normal type in point *3.* of Proposition 4.

(ii) There always exists $\mu' > 0$ such that if $\mu < \mu'$, then $\sigma^{\mu}_N(L)^*$ is the unique root of $\Delta^{\mu}_{N,L}$ in the open interval $(0,1)$. This result follows from the fact that when μ is small enough, $\Delta^{\mu}_{N,L}$ must have a unique root, because of the described characteristics of $\Delta^0_{N,L}$, $\Delta^1_{N,L}$ and $\Delta^{\mu}_{N,L}$. Hence, $\Delta^{\mu}_{N,L} = (1-\mu)\Delta^0_{N,L} + \mu\Delta^1_{N,L} < 0$ for any $\sigma_N(L) < \sigma^{\mu}_N(L)^*$ and $\Delta^{\mu}_{N,L} = (1-\mu)\Delta^0_{N,L} + \mu\Delta^1_{N,L} > 0$ for any $\sigma_N(L) > \sigma^{\mu}_N(L)^*$. See Fig. 2. This implies that in the equilibrium, we necessarily have $0 < \sigma^{\mu}_N(L)^* < \sigma^0_N(L)^*$. This result describes the equilibrium behavior of the normal type in point *1.* of Proposition 4. This equilibrium strategy of the normal type is unique. Note also that in this point of the Proposition, $\sigma^{\mu}_N(L)^* = \sigma'$, with $\sigma' = \sigma^{\mu}_N(L)^{Sup*}$, as σ' is the unique (then, the highest) equilibrium probability with which the normal type takes action L after signal L. Then, by Lemma 9, σ' decreases in μ.

(iii) In addition, since $(\sigma^{\mu}_N(L)^*, \sigma^{\mu}_N(R)^*) = (\sigma'', 0)$ is the most informative equilibrium strategy of the normal type, then $\sigma'' = \sigma^{\mu}_N(L)^{Sup*}$. Then, by Lemma 9, σ'' decreases in μ, and $\sigma'' < \sigma'$.

Next, we prove these lemmas.

Lemma 7. *Consider $\mu = 0$. Let $\alpha''_B = \frac{2+\alpha_W - \sqrt{9\alpha_W^2 - 4\alpha_W + 4}}{4} < \frac{1}{2}$.*

1. If $\alpha_B < \alpha''_B$, then in the unique equilibrium, $(\sigma^0_N(L)^*, \sigma^0_N(R)^*) = \left(1 - \frac{2\alpha_B\alpha_W + \alpha_B\alpha_N}{(\alpha_W - \alpha_B)\alpha_N}, 0\right)$.

2. If $\alpha_B \geq \alpha''_B$, then $(\sigma^0_N(L)^*, \sigma^0_N(R)^*) = (0,0)$ is the unique equilibrium strategy of the normal type.

Proof. The following results prove the lemma (they are proven below).

(1) $\Delta^0_{N,L}\Big|_{\sigma_N(L)=1} > 0$.

(2) $\Delta^0_{N,L}\Big|_{\sigma_N(L)=0} < 0 \iff \alpha_B < \alpha''_B = \frac{2+\alpha_W - \sqrt{9\alpha_W^2 - 4\alpha_W + 4}}{4}$, where $\alpha''_B < \frac{1}{2}\alpha_W$; hence, $\alpha''_B < \frac{1}{2}$.

(3) $\frac{\partial \Delta^0_{N,L}}{\partial \sigma_N(L)} > 0$ if $\alpha_B < \frac{1}{2}\alpha_W$.

(4) $\Delta^0_{N,L} > 0$ if $\alpha_B > \frac{1}{2}\alpha_W$.

(5) $\Delta^0_{N,L} = 0 \iff \sigma_N(L) = 1 - \frac{2\alpha_B\alpha_W + \alpha_B\alpha_N}{(\alpha_W - \alpha_B)\alpha_N}$.

First, note that result (3) implies that $\Delta^0_{N,L}$ is increasing in $\sigma_N(L)$ when $\alpha_B < \alpha''_B$. In addition, $\Delta^0_{N,L}\Big|_{\sigma_N(L)=1} > 0$ and $\Delta^0_{N,L}\Big|_{\sigma_N(L)=0} < 0$ if and only if $\alpha_B < \alpha''_B$. Therefore, if $\alpha_B \in (0, \alpha''_B)$, there is a unique equilibrium, which is $\Delta^0_{N,L}(\sigma^0_N(L)^*) = 0$, with $\sigma^0_N(L)^* = 1 - \frac{2\alpha_B\alpha_W + \alpha_B\alpha_N}{(\alpha_W - \alpha_B)\alpha_N}$. Second, note also that if $\alpha_B \in \left(\alpha''_B, \frac{1}{2}\alpha_W\right)$, then $\Delta^0_{N,L} > 0$ by results (2) and (3), and if $\alpha_B \in \left(\frac{1}{2}\alpha_W, 1\right)$, then $\Delta^0_{N,L} > 0$ by result (4). This second case implies that the optimal strategy of the normal type is $(\sigma^0_N(L)^*, \sigma^0_N(R)^*) = (0,0)$ for any $\alpha_B > \alpha''_B$. Now, by Lemma 3, we know that, in this case, the wise type could find it optimal to use a strategy that does not always follow the agent's signal. In this case, Proposition 8 guarantees that the normal type has a unique equilibrium strategy, which is to always take action R.

Next, we prove the results stated at the beginning of the lemma and that allowed us to prove this lemma.

Proof of result *(1)*: from Eqs. (4) and (5) and the beliefs in Tables 1–2, when $\eta = \beta = 1$, we have $\Delta^0_{N,L}\Big|_{\sigma_N(L)=1} = \frac{2\alpha_W + \alpha_N}{\alpha_W + \alpha_N} - \frac{2\alpha_W + \alpha_N}{2\alpha_B + \alpha_W + \alpha_N} > 0$.

Proof of result *(2)*: $\Delta^0_{N,L}\Big|_{\sigma_N(L)=0} = \frac{\alpha_W}{\alpha_W + 2\alpha_N} + \frac{2\alpha_B - \alpha_W}{2\alpha_B + \alpha_W} = \frac{\alpha_W}{\alpha_W + 2(1-\alpha_B-\alpha_W)} + \frac{2\alpha_B - \alpha_W}{2\alpha_B + \alpha_W} > 0 \iff -2\alpha_B^2 + \alpha_B\alpha_W + 2\alpha_B + \alpha_W^2 - \alpha_W > 0 \iff \alpha_B > \alpha''_B = \frac{1}{4}\left(2 + \alpha_W - \sqrt{9\alpha_W^2 - 4\alpha_W + 4}\right)$. Note that $\alpha''_B = \frac{1}{4}\left(2 + \alpha_W - \sqrt{9\alpha_W^2 - 4\alpha_W + 4}\right) < \frac{1}{2}\alpha_W \iff 2 - \alpha_W < \sqrt{9\alpha_W^2 - 4\alpha_W + 4}$, and $\sqrt{9\alpha_W^2 - 4\alpha_W + 4} > \sqrt{\alpha_W^2 - 4\alpha_W + 4} = 2 - \alpha_W$.

Proof of result *(3)*: $\frac{\partial \Delta^{\mu}_{N,L}}{\partial \sigma_N(L)} = \frac{\alpha_W\alpha_N}{(\alpha_W + 2\alpha_N - \alpha_N\sigma_N(L))^2} + \frac{(\alpha_W - 2\alpha_B)\alpha_N}{(2\alpha_B + \alpha_W + \alpha_N\sigma_N(L))^2} > 0$ when $\alpha_B < \frac{1}{2}\alpha_W$.

Proof of result *(4)*: $\Delta^0_{N,L} = \frac{\alpha_W}{\alpha_W + (2-\sigma_N(L))\alpha_N} + \frac{2\alpha_B - \alpha_W}{2\alpha_B + \alpha_W + \sigma_N(L)\alpha_N} > 0$.

Proof of result *(5)*: $\Delta^0_{N,L} = \frac{\alpha_W}{\alpha_W + (2-\sigma_N(L))\alpha_N} + \frac{2\alpha_B - \alpha_W}{2\alpha_B + \alpha_W + \sigma_N(L)\alpha_N} = 0 \iff \sigma_N(L) = 1 - \frac{2\alpha_B\alpha_W + \alpha_B\alpha_N}{(\alpha_W - \alpha_B)\alpha_N}$. ♦

Lemma 8. *Consider* $\mu = 1$ *and* $\alpha''_B = \frac{2+\alpha_W-\sqrt{9\alpha_W^2-4\alpha_W+4}}{4}$. *For any* $\alpha_W \in (0,1)$, *there always exists* $\alpha'_B = \frac{\alpha_W\left(2+\alpha_W+\alpha_W^2-2\sqrt{2}\sqrt{\alpha_W(\alpha_W+1)}\right)}{(\alpha_W+2)^2} \in (0,1)$, *with* $\alpha'_B < \alpha''_B$, *such that for any* $\alpha_B \in (\alpha'_B, \alpha''_B)$, *there exists* $\gamma' = \frac{\left(\alpha_B^2-\alpha_W^2\right)}{2\sqrt{2\alpha_B\alpha_W}-(2\alpha_B+\alpha_W)(1-\alpha_B+\alpha_W)} > \frac{1}{2}$. *Then, for any* $\alpha_B \in (\alpha'_B, \alpha''_B)$ *and* $\gamma \in (\frac{1}{2}, \gamma')$, *we have* $\Delta^1_{N,L} > 0$. *Hence, the unique equilibrium strategy of the normal type is* $(\sigma^1_N(L)^*, \sigma^1_N(R)^*) = (0,0)$.

Proof. From Eq. (4) and the beliefs in Tables 1–2, we have the following:

$$\Delta^1_{N,L} = \frac{(1-\gamma)\alpha_W}{\alpha_W + (\gamma+(1-\gamma)(1-\sigma_N(L)))\alpha_N} - \frac{\gamma(\alpha_W-\alpha_B)}{\alpha_B+\alpha_W+\gamma\sigma_N(L)\alpha_N} + \frac{(1-\gamma)\alpha_B}{\alpha_B+(1-\gamma)\sigma_N(L)\alpha_N}.$$

Note that if $\alpha_B > \alpha_W$, the expression above is positive, in which case $\sigma^1_N(L)^* = 0$. Hence, hereafter, we focus on $\alpha_B < \alpha_W$. Note that $\alpha''_B < \alpha_W$ as shown above.

For the case $\alpha_B < \alpha_W$, it is easy to obtain that $\Delta^1_{N,L} > 0$ if and only if the following second-degree polynomial $p(1-\sigma_N(L))$ is positive:

$$\begin{aligned} p(\cdot) = {} & (1-\sigma_N(L))^2 2\gamma\alpha_N^2(1-\gamma)^2\left(\alpha_W-\alpha_B\right) \\ & + (1-\sigma_N(L))\left(\alpha_N(1-\gamma)(-4\gamma^2\alpha_N\alpha_B+4\gamma^2\alpha_N\alpha_W-4\gamma\alpha_B\alpha_W+2\gamma\alpha_N\alpha_B+2\gamma\alpha_W^2-3\gamma\alpha_N\alpha_W+\alpha_B^2-\alpha_W^2)\right) \\ & + \gamma(\alpha_W+\gamma\alpha_N)\left(\alpha_B-\alpha_N(\gamma-1)\right)(\alpha_B-\alpha_W) - \alpha_W(\gamma-1)\left(\alpha_B-\alpha_N(\gamma-1)\right)(\alpha_B+\alpha_W+\gamma\alpha_N) \\ & - \alpha_B(\alpha_W+\gamma\alpha_N)(\gamma-1)(\alpha_B+\alpha_W+\gamma\alpha_N). \end{aligned} \tag{11}$$

To facilitate the analysis, in the following, we denote $(1-\sigma_N(L))$ by z and refer to this polynomial, Eq. (11), as $p(z) = a(z)^2+b(z)+c$. Note that if $p(z) > 0$, then $\Delta^1_{N,L} > 0$. The first derivative is $p'(z) = 2a(z)+b$, and the second one is $p''(z) = 2a$. Additionally, $p(z)$ is convex in z when $\alpha_W > \alpha_B$, as in this case $a = 2\gamma\alpha_N^2(1-\gamma)^2\left(\alpha_W-\alpha_B\right) > 0$. The first derivative provides the minimum value of $p(z)$: $p'(z_{\min}) = 2az_{\min}+b = 0 \iff z_{\min} = \frac{-b}{2a}$. Consequently, if $p(z_{\min}) > 0$, then $p(z) > 0$, and therefore, $\Delta^1_{N,L} > 0$.

Next, we determine the conditions for $p(z_{\min}) > 0$. Note that $p(z_{\min}) = a(\frac{-b}{2a})^2 + b(\frac{-b}{2a}) + c = c - \frac{b^2}{4a}$. From Eq. (11), we can obtain the value of $c-\frac{b^2}{4a}$, which is positive if the next polynomial in γ is positive:

$$\begin{aligned} pol(\gamma) = {} & \gamma^2(8\alpha_B^3\alpha_W-4\alpha_B^2\alpha_N^2-8\alpha_B\alpha_W^3+4\alpha_B\alpha_W\alpha_N^2-4\alpha_W^4-4\alpha_W^3\alpha_N-\alpha_W^2\alpha_N^2) \\ & + \gamma\left(2(\alpha_W-\alpha_B)(2\alpha_B+\alpha_W)(\alpha_B+\alpha_W)(2\alpha_W+\alpha_N)\right) \\ & - (\alpha_B^2-\alpha_W^2)^2 \end{aligned} \tag{12}$$

Note that $pol(\gamma) > 0 \iff p(z_{\min}) > 0 \Rightarrow p(z) > 0 \iff \Delta^1_{N,L} > 0$. Then, we determine the conditions for $pol(\gamma) > 0$. We can easily check that under condition $\alpha_B < \alpha_W$:

(1) $pol(\gamma)$ is a concave function in γ.
(2) $\frac{dpol(\gamma)}{d\gamma}\Big|_{\gamma=\frac{1}{2}} < 0$

Now, since $pol(\gamma)$ is concave in $\gamma \in (\frac{1}{2}, 1)$ and decreasing in $\gamma = \frac{1}{2}$, polynomial $pol(\gamma)$ is necessarily decreasing in $\gamma \in (\frac{1}{2}, 1)$. Hence, there are only two possibilities: (i) $pol(\gamma)$ is negative for all $\gamma \in \left(\frac{1}{2}, 1\right)$, and (ii) $pol(\gamma)$ is positive for all $\gamma \in (\frac{1}{2}, \gamma')$ and negative for all $\gamma \in (\tilde{\gamma}, 1)$, where γ' is the greatest real root of $pol(\gamma) = 0$. If $pol(\gamma = \frac{1}{2}) > 0$, we are in the second case.

Now, from Eq. (12) and setting $\alpha_N = 1-\alpha_W-\alpha_B$, we obtain the following:

$$pol(\gamma = \tfrac{1}{2}) = \left(-\frac{1}{4}\alpha_W^2-\alpha_W-1\right)\alpha_B^2 + \left(\frac{1}{2}\alpha_W^3+\frac{1}{2}\alpha_W^2+\alpha_W\right)\alpha_B + \left(\frac{1}{2}\alpha_W^3-\frac{1}{4}\alpha_W^4-\frac{1}{4}\alpha_W^2\right).$$

The expression above is concave in α_B and has the following two real roots:

$$\alpha'_B = \frac{\left(2\alpha_W+\alpha_W^2+\alpha_W^3-2\sqrt{2}\sqrt{\alpha_W^4+\alpha_W^3}\right)}{(\alpha_W+2)^2} \text{ and } \alpha_B^{*(+)} = \frac{\left(2\alpha_W+\alpha_W^2+\alpha_W^3+2\sqrt{2}\sqrt{\alpha_W^4+\alpha_W^3}\right)}{(\alpha_W+2)^2}.$$

Consequently, if $\alpha_B \in \left(\alpha'_B, \alpha_B^{*(+)}\right)$, $pol(\gamma = \frac{1}{2}) > 0$, and then, $pol(\gamma)$ is positive for all $\gamma \in (\frac{1}{2}, \gamma')$ and negative for all $\gamma \in (\gamma', 1)$, where γ' is the greatest real root of $pol(\gamma)$, i.e., $\gamma' = \frac{\left(\alpha_B^2-\alpha_W^2\right)}{2\sqrt{2\alpha_B\alpha_W}-(2\alpha_B+\alpha_W)(1-\alpha_B+\alpha_W)}$.

To conclude the proof of Lemma 8, we need to show that for any α_W, $\alpha'_B < \alpha''_B < \alpha_B^{*(+)}$. To prove this, note that after some algebra, we obtain $\alpha''_B < \alpha_B^{*(+)} \iff \frac{9\alpha_W^6+68\alpha_W^5+184\alpha_W^4+208\alpha_W^3+64\alpha_W^2}{(\alpha_W+2)^4} > 0$, which always holds. We also obtain $\alpha'_B < \alpha''_B \iff 3\alpha_W+3\alpha_W^2-2 < \sqrt{8\alpha_W^2+8\alpha_W}$. Note that the expression on the left-hand side of the inequality is a convex function in α_W, whereas the expression on the right-hand side is concave in α_W. Additionally, the value of the left-hand side expression evaluated at $\alpha_W = 0$ is smaller than the value of the right-hand side expression evaluated at $\alpha_W = 0$, and both are equal at $\alpha_W = 1$. Now, since $\alpha_W \in (0,1)$, the inequality holds.

In summary, if $\mu = 1$, $\alpha_B \in \left(\alpha'_B, \alpha''_B\right)$ and $\gamma \in (\frac{1}{2}, \gamma')$, then $pol(\gamma) > 0$. Hence, $p(z) = p(1 - \sigma_N(L)) > 0$ and $\Delta^1_{N,L} > 0$. Consequently, in equilibrium, $\sigma^1_N(L)^* = 0$. Finally, by Lemma 3 and Proposition 8, the proof follows. This concludes the proof of Lemma 8. ♦

Lemma 9. *$\sigma^\mu_N(L)^{Sup*}$ decreases as μ increases.*

Proof. We first define $\sigma^\mu_N(L)^{Sup*}$ as the highest equilibrium probability with which the normal type takes action L after signal L. We refer to $\sigma^\mu_N(L)^{Sup*}$ as the supremum (of all $\sigma^\mu_N(L)^*$).

Now, note that if $\alpha_B \in (\alpha'_B, \alpha''_B)$ and $\gamma \in (\frac{1}{2}, \gamma')$, then $\Delta^1_{N,L} > 0$ for all $\sigma_N(L) \in (0,1)$. In addition, $\Delta^0_{N,L} < 0$ for all $\sigma_N(L) \in (0, \sigma^0_N(L)^*)$, and $\Delta^0_{N,L} > 0$ for all $\sigma_N(L) \in (\sigma^0_N(L)^*, 1)$ (see Lemmas 7 and 8). Therefore, $\Delta^\mu_{N,L} = (1-\mu)\Delta^0_{N,L} + \mu\Delta^1_{N,L} > 0$ for all $\sigma_N(L) \in (\sigma^0_N(L)^*, 1)$, which implies $\sigma^\mu_N(L)^* < \sigma^0_N(L)^*$ for any $\mu > 0$, hence $\sigma^\mu_N(L)^{Sup*} < \sigma^0_N(L)^*$ for any $\mu > 0$. Then, $\sigma^\mu_N(L)^{Sup*} \in (0, \sigma^0_N(L)^*)$. In addition, $\Delta^\mu_{N,L} > 0$ necessarily for all $\sigma^\mu_N(L) \in (\sigma^\mu_N(L)^{Sup*}, \sigma^0_N(L)^*)$.

Now, note that $\frac{\partial \Delta^\mu_{N,L}}{\partial \mu} = -\Delta^0_{N,L} + \Delta^1_{N,L} > 0$ for all $\sigma_N(L) \in (0, \sigma^0_N(L)^*)$ since $\Delta^1_{N,L} > 0$ for all $\sigma_N(L) \in (0,1)$ and $\Delta^0_{N,L} < 0$ for all $\sigma_N(L) \in (0, \sigma^0_N(L)^*)$.

Therefore, if μ increases, $\Delta^\mu_{N,L}$ increases, and as $\Delta^\mu_{N,L} > 0$ for all $\sigma_N(L) \in (\sigma^\mu_N(L)^{Sup*}, \sigma^0_N(L)^*)$, $\sigma^\mu_N(L)^{Sup*}$ has to decrease with μ. ♦

Lemma 10. $\alpha''_B < \widetilde{\alpha}_B$.

Proof. From the proof of Lemma 3, we know that when the normal type uses a non-informative strategy, $\frac{\partial \Delta^\mu_{W,L}}{\partial \sigma_W(L)} < 0$, and $\Delta^\mu_{W,L}\Big|_{\sigma_W(L)=0} > 0$. Additionally, in this case, $\Delta^\mu_{W,L}\Big|_{\sigma_W(L)=1} \geq 0 \iff \alpha_B \geq \widetilde{\alpha}_B$.

Also, from expression (8), when $f\left(\lambda_W(a,X), \lambda_{\bar{B}}(a,X)\right) = \lambda_W(a,X) + \lambda_{\bar{B}}(a,X)$, we have the following:

$$\begin{aligned}\Delta^\mu_{W,L}\Big|_{\sigma_W(L)=1} &= (1-\mu)\left(\tfrac{2(1-\alpha_B)}{2-2\alpha_B-\alpha_W} - \tfrac{2\alpha_W}{\alpha_W+2\alpha_B}\right) + \mu\left(\tfrac{\alpha_B-\alpha_W}{\alpha_B+\alpha_W}\right)\\ &= -\mu\left(\tfrac{\alpha_W-\alpha_B}{\alpha_B+\alpha_W}\right) - (1-\mu)\left(\tfrac{2\alpha_W}{\alpha_W+2\alpha_B} - \tfrac{2(1-\alpha_B)}{2-2\alpha_B-\alpha_W}\right),\end{aligned}$$

where

$\frac{2\alpha_W}{\alpha_W+2\alpha_B} - \frac{2(1-\alpha_B)}{2-2\alpha_B-\alpha_W} > 0 \iff \alpha_B \leq \frac{2+\alpha_W-\sqrt{9\alpha_W^2-4\alpha_W+4}}{4} = \alpha''_B$. In addition, $\alpha''_B < \alpha_W$; hence, $-\mu\left(\frac{\alpha_W-\alpha_B}{\alpha_B+\alpha_W}\right) < 0$ for any $\alpha_B \leq \alpha''_B$.

Consequently, if $\alpha_B \leq \alpha''_B$, then $\Delta^\mu_{W,L}\Big|_{\sigma_W(L)=1} < 0$, which implies $\alpha''_B < \widetilde{\alpha}_B$. ♦

This completes the proof of Proposition 4. □

Proof of Proposition 5. By Proposition 2 and Definition 3, $\sigma^\mu_N(R)^* = 0\ \forall \mu \in [0,1]$.

By Corollary 1, if the wise type does not use the honest strategy, then $(\sigma^\mu_N(L)^*, \sigma^\mu_N(R)^*) = (0,0)$ for all μ. Hence, if we want to analyze whether an informative equilibrium exists, we have to consider that the wise type uses the honest strategy, i.e., $(\sigma^\mu_W(L)^*, \sigma^\mu_W(R)^*) = (1,0)$. This is what we do in point *1.* of Proposition 5. It is also what we assume in order to prove point *2.* of this proposition. The difference is that, in the latter case, we obtain that the equilibrium strategy of the normal type is $(\sigma^\mu_N(L)^*, \sigma^\mu_N(R)^*) = (0,0)$, which is non-informative. Then, by Proposition 1, in equilibrium either $(\sigma^\mu_W(L)^*, \sigma^\mu_W(R)^*) = (0,0)$, $(\sigma^\mu_W(L)^*, \sigma^\mu_W(R)^*) = (1,0)$ or $(\sigma^\mu_W(L)^*, \sigma^\mu_W(R)^*) = (x_1, 0)$, with $x_1 \in (0,1)$.

To prove Proposition 5 we analyze the expression $\Delta^\mu_{N,L} = (1-\mu)\Delta^0_{N,L} + \mu\Delta^1_{N,L}$ in the two extreme cases $\mu = 0$ and $\mu = 1$. We consider expression (4), with $f\left(\lambda_W(a,X), \lambda_{\bar{B}}(a,X)\right) = \eta\lambda_W(a,X) + \lambda_{\bar{B}}(a,X)$.

Let us first define some thresholds:

$$\bar{\alpha}_B = 2\sqrt{2}\sqrt{\alpha_W^2 + \alpha_W} - 2\alpha_W,$$

$$\bar{\gamma} = \frac{1}{2}\frac{\alpha_B + \alpha_W}{2\alpha_B + \alpha_W}\left(1 + \sqrt{2 + \frac{2}{(2\alpha_B + \alpha_W - 2)}}\right),$$

$$\bar{\eta}' = \frac{\alpha_B(2\alpha_B + \alpha_W - 2)}{\alpha_W(2\alpha_B + \alpha_W - 1)},$$

$$\bar{\eta}'' = \frac{\alpha_B\left(2\sqrt{2}\gamma\sqrt{(\gamma(2\alpha_B+\alpha_W-1)-\alpha_B-\alpha_W)(\alpha_B(2\gamma-1)+\alpha_W(\gamma-1))} + (\gamma(2\alpha_B+\alpha_W-1)-\alpha_B-\alpha_W)(\alpha_B(2\gamma-1)+\alpha_W(\gamma-1)) + 2\gamma^2\right)}{\alpha_W(\alpha_B(2\gamma-1)+\alpha_W(\gamma-1)+\gamma)^2}.$$

.

Next, Lemma 11 analyzes the behavior of $\Delta^\mu_{N,L}$ when $\mu = 0$, and Lemma 12 does it when $\mu = 1$.

Lemma 11. *Consider $\sigma_N(L) = 0$. Then, $\Delta^0_{N,L} < 0$ if $\alpha_B < \frac{1-\alpha_W}{2}$ and $\eta > \bar{\eta}'$.*

Proof. From (4), we have:

$$\Delta^0_{N,L}\Big|_{\sigma_N(L)=0} < 0 \iff \frac{2\alpha_B(2\alpha_B+\alpha_W-2)-2\alpha_W\eta(2\alpha_B+\alpha_W-1)}{(2\alpha_B+\alpha_W-2)(2\alpha_B+\alpha_W)} < 0.$$

Simple algebra shows that this expression is smaller than zero when $\alpha_B < \frac{1-\alpha_W}{2}$ and $\eta > \frac{\alpha_B(2\alpha_B+\alpha_W-2)}{\alpha_W(2\alpha_B+\alpha_W-1)}$, with $\bar{\eta}' = \frac{\alpha_B(2\alpha_B+\alpha_W-2)}{\alpha_W(2\alpha_B+\alpha_W-1)}$. ♦

Lemma 12. *$\Delta^1_{N,L} > 0$ if $\alpha_B < \bar{\alpha}_B$, $\gamma < \bar{\gamma}$ and $\eta \in (\bar{\eta}', \bar{\eta}'')$, with $0 < \bar{\eta}' < \bar{\eta}'' < 1$.*

Proof. From (4), we have:

$\Delta^1_{N,L} > 0 \iff \frac{\alpha_W(\gamma-1)\eta}{(\gamma-1)\sigma_N(L)(\alpha_B+\alpha_W-1)+\alpha_B-1} + \frac{\gamma(\alpha_B-\alpha_W\eta)}{\alpha_B+\alpha_W-\gamma\sigma_N(L)(\alpha_B+\alpha_W-1)} + \frac{\alpha_B-\alpha_B\gamma}{(\gamma-1)\sigma_N(L)(\alpha_B+\alpha_W-1)+\alpha_B} > 0.$

This expression can be written as $\frac{Nu}{De}$, with:

$Nu = -2(1-\gamma)^2\gamma(\alpha_B+\alpha_W-1)^2(\alpha_B-\alpha_W\eta)\sigma_N(L)^2 - (\gamma-1)\sigma_N(L)(\alpha_B+\alpha_W-1)(\alpha_W\eta(-\alpha_B(2\gamma+1)+\alpha_W(\gamma-1)+\gamma)+\alpha_B(2\alpha_B\gamma+\alpha_B-(\alpha_W+2)\gamma+\alpha_W)) + \alpha_B\alpha_W\eta(\alpha_B-(\alpha_W+1)\gamma+\alpha_W) - (\alpha_B-1)\alpha_B(\alpha_B-\alpha_W\gamma+\alpha_W),$

$De = ((\gamma-1)\sigma_N(L)(\alpha_B+\alpha_W-1)+\alpha_B-1)((\gamma-1)\sigma_N(L)(\alpha_B+\alpha_W-1)+\alpha_B)(\gamma\sigma_N(L)(\alpha_B+\alpha_W-1)-\alpha_B-\alpha_W).$

It can be shown that the denominator is always positive. Hence, if the numerator is positive, $\Delta^1_{N,L} > 0$.

We note that when $\alpha_B < \frac{1-\alpha_W}{2}$ and $\eta > \bar{\eta}'$, $Nu(\sigma_N(L))$ is a convex function in $\sigma_N(L)$. Thus, we next calculate the minimum of $Nu(\sigma_N(L))$, that we denote by $\sigma^M_N(L)$. The value of this function evaluated at the minimum is $Nu(\sigma^M_N(L)) = \frac{Nu'}{De'}$, with:

$$Nu' = -2\alpha_B\alpha_W\eta\left(\gamma^2\left(4\alpha_B^2+\alpha_B(4\alpha_W-2)+(\alpha_W-1)\alpha_W+2\right)-\gamma(4\alpha_B+2\alpha_W-1)(\alpha_B+\alpha_W)+(\alpha_B+\alpha_W)^2\right)+\alpha_B^2(-\gamma(2\alpha_B+\alpha_W-2)+\alpha_B+\alpha_W)^2+\alpha_W^2\eta^2(\alpha_B(2\gamma-1)+\alpha_W(\gamma-1)+\gamma)^2,$$

$$De' = 8\gamma(\alpha_B-\alpha_W\eta).$$

Note that if $Nu(\sigma^M_N(L)) \geq 0$, then $Nu(\sigma_N(L)) > 0$ for any $\sigma_N(L) \in (0,1)$, because of the convexity of $Nu(\sigma_N(L))$. It can be shown that, on the one hand, $Nu(\sigma^M_N(L)) = 0$ when $\eta = \bar{\eta}''$, with $\bar{\eta}''$ as defined above. On the other hand, it can be shown that $Nu(\sigma^M_N(L))$ is decreasing in η when $\alpha_B < \frac{1-\alpha_W}{2}$ and $\eta > \bar{\eta}'$. In addition, if $\gamma < \bar{\gamma}$ and $\alpha_B < \bar{\alpha}_B$, with $\bar{\alpha}_B = 2\sqrt{2}\sqrt{\alpha_W^2+\alpha_W} - 2\alpha_W$, then $\bar{\eta}' < \bar{\eta}''$. Note that $\bar{\alpha}_B < \frac{1-\alpha_W}{2}$. Concluding, if $\alpha_B < 2\sqrt{2}\sqrt{\alpha_W^2+\alpha_W} - 2\alpha_W$, $\gamma < \bar{\gamma}$, and $\eta \in (\bar{\eta}', \bar{\eta}'')$, then $Nu(\sigma^M_N(L)) > 0$, which implies $Nu(\sigma_N(L)) > 0$ and $\Delta^1_{N,L} > 0$. ♦

Now, by the continuity of function $\Delta^\mu_{N,L} = (1-\mu)\Delta^0_{N,L} + \mu\Delta^1_{N,L}$, if $\Delta^0_{N,L}\Big|_{\sigma_N(L)=0} < 0$, there exists $\bar{\mu}' > 0$ such that for all $\mu \in (0, \bar{\mu}')$, $\Delta^\mu_{N,L}\Big|_{\sigma_N(L)=0} < 0$, which implies $\sigma^\mu_N(L)^* > 0$ for all $\mu < \bar{\mu}'$. Analogously, if $\Delta^1_{N,L} > 0$, then there exists $\bar{\mu}'' > 0$ such that if $\mu \in (\bar{\mu}'', 1)$, then $\Delta^\mu_{N,L} > 0$, which implies $\sigma^\mu_N(L)^* = 0$ for all $\mu > \bar{\mu}''$.

To complete the proof, Lemma 13 below shows that $2\sqrt{2}\sqrt{\alpha_W^2+\alpha_W} - 2\alpha_W < \widetilde{\alpha}_B$, which implies that if $\sigma^\mu_N(L)^* = 0$ and $\alpha_B < \bar{\alpha}_B$, then the equilibrium behavior of the wise type is described by point *2.b* of Proposition 1.

Lemma 13. $2\sqrt{2}\sqrt{\alpha_W^2+\alpha_W} - 2\alpha_W < \widetilde{\alpha}_B$.

Proof. From the proof of Lemma 3, we know that when the normal type uses a non-informative strategy, $\frac{\partial \Delta^\mu_{W,L}}{\partial \sigma_W(L)} < 0$, and $\Delta^\mu_{W,L}\Big|_{\sigma_W(L)=0} > 0$. Additionally, in this case, $\Delta^\mu_{W,L}\Big|_{\sigma_W(L)=1} \geq 0 \iff \alpha_B \geq \widetilde{\alpha}_B$.

Also, from expression (8), when $f\left(\lambda_W(a,X), \lambda_{\bar{B}}(a,X)\right) = \eta\lambda_W(a,X) + \lambda_{\bar{B}}(a,X)$, we have the following:

$$\Delta^\mu_{W,L}\Big|_{\sigma_W(L)=1} = (1-\mu)\left(\frac{\alpha_W\eta}{2-2\alpha_B-\alpha_W} - \frac{\alpha_W\eta}{2\alpha_B+\alpha_W} + \frac{2\alpha_B}{2\alpha_B+\alpha_W}\right) + \mu\left(\frac{\alpha_B-\alpha_W\eta}{\alpha_B+\alpha_W}\right).$$

It can be shown that:

(1) $\frac{\alpha_W\eta}{2-2\alpha_B-\alpha_W} - \frac{\alpha_W\eta}{2\alpha_B+\alpha_W} + \frac{2\alpha_B}{2\alpha_B+\alpha_W} < 0$ if and only if $\alpha_B < \frac{1-\alpha_W}{2}$ and $\eta > \frac{\alpha_B(2\alpha_B+\alpha_W-2)}{\alpha_W(2\alpha_B+\alpha_W-1)}$.

(2) $\frac{\alpha_B-\alpha_W\eta}{\alpha_B+\alpha_W} < 0$ if and only if $\alpha_B < \frac{1-\alpha_W}{2}$ and $\eta > \frac{\alpha_B}{\alpha_W}$.

(3) $\frac{\alpha_B}{\alpha_W} < \frac{\alpha_B(2\alpha_B+\alpha_W-2)}{\alpha_W(2\alpha_B+\alpha_W-1)}$ if and only if $\alpha_B < \frac{1-\alpha_W}{2}$.

Therefore, if $\alpha_B < \frac{1-\alpha_W}{2}$ and $\eta > \frac{\alpha_B(2\alpha_B+\alpha_W-2)}{\alpha_W(2\alpha_B+\alpha_W-1)}$, then $\Delta^\mu_{W,L}\Big|_{\sigma_W(L)=1} < 0$.

To conclude, note that $2\sqrt{2}\sqrt{\alpha_W^2+\alpha_W} - 2\alpha_W < \frac{1-\alpha_W}{2}$. ♦

This completes the proof of Proposition 5. □

Proof of Proposition 6. Let $\theta' = \hat{\theta} - \epsilon$ and $\theta'' = \hat{\theta} + \epsilon$, with $\hat{\theta} = 1/2$ and $\epsilon \sim 0$. For a given μ, from the continuity of the payoff function, the result follows. □

Proof of Proposition 7. For all $\mu > 0$ and $s \in \{L, R\}$, now Definition 1 implies $\Delta^{\mu}_{B,s} = \Pi^{\mu}_{N,s}(R) - (\Pi^{\mu}_{N,s}(L) + \phi)$. This can be rewritten as $\Delta^{\mu}_{B,s} = \Delta^{\mu}_{N,s} - \phi$, with $\Delta^{\mu}_{N,s}$ given by expressions (4)–(5). Then, for all $\phi > 0$, $\Delta^{\mu}_{N,s} > \Delta^{\mu}_{B,s}$. Given $s' \in \{L, R\}$, this means that if $\Delta^{\mu}_{N,s'} \leq 0$, then $\Delta^{\mu}_{B,s'} < 0$; hence $\sigma^{\mu}_B(s')^* = 1$. Additionally, if $\Delta^{\mu}_{N,s'} > 0$, then $\Delta^{\mu}_{B,s'} \lesseqgtr 0$. Then, there exists $\hat{\phi} > 0$ such that for all $\phi > \hat{\phi}$, $\Delta^{\mu}_{B,s'} < 0$; hence, $\sigma^{\mu}_B(s')^* = 1$ for all $s' \in \{L, R\}$. □

Proposition 9. *Consider $\mu \in [0, 1]$, $\beta = 0$ and $\eta = 1$, i.e., the objective function of the agent is $\Pi(a, X) = \lambda_W(a, X)$. There exists $\bar{\alpha}^a_B$ and $\bar{\alpha}^b_B$, with $0 < \bar{\alpha}^a_B < \bar{\alpha}^b_B < 1$, such that the following holds:*

1. *For any $\alpha_B \leq \bar{\alpha}^a_B$, there exists $\bar{\mu} > 0$ such that in the unique informative equilibrium, $\sigma^{\mu}_N(L)^* = 1$ when $\mu > \bar{\mu}$, and $\sigma^{\mu}_N(L)^* < 1$ when $\mu < \bar{\mu}$. In the latter case, $\frac{d\sigma^{\mu}_N(L)^*}{d\mu} > 0$. Additionally, if $\alpha_B \to 0$, then $\bar{\mu} \to 0$.*
2. *If $\alpha_B \in (\bar{\alpha}^a_B, \bar{\alpha}^b_B)$, then in the unique informative equilibrium, $\sigma^{\mu}_N(L)^* < 1$ for all μ. In this case, $\frac{d\sigma^{\mu}_N(L)^*}{d\mu} > 0$.*
3. *If $\alpha_B \geq \bar{\alpha}^b_B$, then in equilibrium, $\sigma^{\mu}_N(L)^* = 0$, for all μ.*

Proof. The results of Proposition 1, Corollary 1 and Proposition 2 are derived for any increasing function $f(\lambda_W(a, W), \lambda_{\bar{B}}(a, X))$. Hence, they also apply to this case. Consequently, first note that in any informative equilibria the wise type always follows the honest strategy.[26] Hence, in the proof we consider that the wise type uses the honest strategy and focus on the behavior of the normal type. Second, by Proposition 2, we can consider $\sigma^{\mu}_N(R)^* = 0$ for all $\mu \in [0, 1]$. For the case $\mu = 0$, we use Definition 3.

The proof of Proposition 9 requires three results, which are proven in three lemmas, Lemmas 14–16. Lemma 14 characterizes the unique equilibrium when $\mu = 0$. It defines threshold $\hat{\alpha}_B = \frac{1-\alpha_W}{2}$ and shows that if $\alpha_B < \hat{\alpha}_B$, then $\sigma^0_N(L)^* = 1 - \frac{\alpha_B}{\alpha_N}$, and if $\alpha_B > \hat{\alpha}_B$, then $\sigma^0_N(L)^* = 0$. Lemma 15 characterizes the unique equilibrium when $\mu = 1$. It defines thresholds $\bar{\alpha}^a_B = \frac{(2\gamma-1)(\gamma(1-\alpha_W)+\alpha_W)}{1-(1-\gamma)2\gamma}$ and $\bar{\alpha}^b_B = \gamma - (1-\gamma)\alpha_W$, with $\bar{\alpha}^a_B < \bar{\alpha}^b_B$, and shows that if $\alpha_B \leq \bar{\alpha}^a_B$, then $\sigma^1_N(L)^* = 1$; if $\bar{\alpha}^a_B < \alpha_B < \bar{\alpha}^b_B$, then $0 < \sigma^1_N(L)^* < 1$; and if $\alpha_B \geq \bar{\alpha}^b_B$, then $\sigma^1_N(L)^* = 0$. Finally, Lemma 16 shows that if $\sigma^{\mu}_N(L)^* \in (0, 1)$, then $\sigma^{\mu}_N(L)^*$ is increasing in μ.

We next study the relationship between thresholds $\hat{\alpha}_B$, $\bar{\alpha}^a_B$ and $\bar{\alpha}^b_B$. Note that $\hat{\alpha}_B < \bar{\alpha}^b_B \iff \frac{1}{2} - \frac{1}{2}\alpha_W < \gamma - \alpha_W(1-\gamma)$, which always holds as $\gamma > \frac{1}{2}$. In addition, $\hat{\alpha}_B < \bar{\alpha}^a_B \iff \frac{1}{2} - \frac{1}{2}\alpha_W < \frac{(2\gamma-1)(\gamma+\alpha_W-\gamma\alpha_W)}{\gamma^2+(\gamma-1)^2} \iff \frac{\alpha_W - 2\gamma^2 - 4\gamma\alpha_W + 2\gamma^2\alpha_W + 1}{4\gamma^2 - 4\gamma + 2} < 0 \iff \gamma > 1 - \frac{1}{1-\alpha_W} + \frac{1}{\sqrt{2 - \frac{4\alpha_W}{1+\alpha^2_W}}}$.

Let $\check{\gamma} = 1 - \frac{1}{1-\alpha_W} + \frac{1}{\sqrt{2 - \frac{4\alpha_W}{1+\alpha^2_W}}}$. Then, if $\gamma < \check{\gamma}$, $\bar{\alpha}^a_B < \hat{\alpha}_B < \bar{\alpha}^b_B$, and if $\gamma \geq \check{\gamma}$, $\hat{\alpha}_B \leq \bar{\alpha}^a_B < \bar{\alpha}^b_B$.

Now, there are two cases to consider. First, suppose $\gamma < \check{\gamma}$, then $\bar{\alpha}^a_B < \hat{\alpha}_B < \bar{\alpha}^b_B$. In the following lemmas, we also prove that $\frac{\partial \Delta^0_{N,L}}{\partial \sigma_N(L)} > 0$, $\frac{\partial \Delta^1_{N,L}}{\partial \sigma_N(L)} > 0$, and $\frac{\partial \Delta^{\mu}_{N,L}}{\partial \sigma_N(L)} > 0$. In this case, we have the following:

(1) If $\alpha_B < \bar{\alpha}^a_B$, then $0 < \sigma^0_N(L)^* < 1$ and $\sigma^1_N(L)^* = 1$. Thus, $\sigma^{\mu}_N(L)^* \in (0, 1]$. Since $\frac{d\sigma^{\mu}_N(L)^*}{d\mu} > 0$ for any $\alpha_B \in (0, \bar{\alpha}^a_B)$, there always exists $\bar{\mu} > 0$ such that $\sigma^{\mu}_N(L)^* < 1$ if $\mu < \bar{\mu}$, and $\sigma^{\mu}_N(L)^* = 1$ if $\mu > \bar{\mu}$. Additionally, note that by Lemma 14, $\sigma^0_N(L)^* = 1 - \frac{\alpha_B}{\alpha_N}$; hence, if $\alpha_B \to 0$, then $\sigma^0_N(L)^* \to 1$, which implies that $\bar{\mu} \to 0$.

(2) If $\alpha_B \in (\bar{\alpha}^a_B, \hat{\alpha}_B)$, then $0 < \sigma^0_N(L)^* < 1$ and $\sigma^1_N(L)^* < 1$. Thus, $\sigma^{\mu}_N(L)^* \in (0, 1)\ \forall \mu$, with $\frac{d\sigma^{\mu}_N(L)^*}{d\mu} > 0$.

(3) If $\alpha_B \in (\hat{\alpha}_B, \bar{\alpha}^b_B)$, then $\sigma^0_N(L)^* = 0$ and $\sigma^1_N(L)^* < 1$. Thus, $\sigma^{\mu}_N(L)^* \in [0, 1)\ \forall \mu$, with $\frac{d\sigma^{\mu}_N(L)^*}{d\mu} > 0$.

(4) If $\alpha_B \geq \bar{\alpha}^b_B$, then $\sigma^0_N(L)^* = 0$ and $\sigma^1_N(L)^* = 0$. Thus, $\sigma^{\mu}_N(L)^* = 0\ \forall \mu$.

The analysis for $\gamma > \check{\gamma}$, then $\hat{\alpha}_B < \bar{\alpha}^a_B < \bar{\alpha}^b_B$, is analogous, and thus, it is omitted.

Next, we prove Lemmas 14–16.

Lemma 14. *Suppose $\mu = 0$. Let $\hat{\alpha}_B = \frac{1-\alpha_W}{2}$.*

(i) If $\alpha_B < \hat{\alpha}_B$, there is a unique equilibrium. In the equilibrium, $\sigma^0_N(L)^* = 1 - \frac{\alpha_B}{\alpha_N}$.

(ii) If $\alpha_B \geq \hat{\alpha}_B$, then $\sigma^0_N(L)^* = 0$ in the unique equilibrium strategy.

Proof. In this case, expressions (4) and (5) simplify to the following:

$$\Delta^0_{N,L} = \Delta^0_{N,R} = \lambda_W(R, 0) - \lambda_W(L, 0) = \frac{\alpha_W}{\alpha_W + (2 - \sigma_N(L))\alpha_N} - \frac{\alpha_W}{2\alpha_B + \alpha_W + \sigma_N(L)\alpha_N}. \tag{13}$$

After some algebra, we obtain the following:

$$\Delta^0_{N,L} = \Delta^0_{N,R} \geq 0 \iff \frac{\alpha_B}{\alpha_N} \geq 1 - \sigma_N(L).$$

[26] Therefore, in points *1.* and *2.* of Proposition 9 we can consider that the wise type follows the honest strategy and focus on the normal type behavior. In point *3.* the normal type plays $(\sigma^{\mu}_N(L)^*, \sigma^{\mu}_N(R)^*) = (0, 0)$. In that case, if $\alpha_B < \widetilde{\alpha}$, by Proposition 1, the honest strategy of the wise type is an equilibrium strategy. When $\alpha_B > \widetilde{\alpha}$, then the unique equilibrium strategy of the wise type is $(\sigma^{\mu}_W(L)^*, \sigma^{\mu}_W(R)^*) = (0, 0)$. Nevertheless, note that in any case the normal type always plays $(\sigma^{\mu}_N(L)^*, \sigma^{\mu}_N(R)^*) = (0, 0)$.

There are two cases:

(i) If $\alpha_B < \alpha_N$, then $\frac{\alpha_B}{\alpha_N} < 1$, and consequently, $\Delta^0_{N,L} = \Delta^0_{N,R} = 0$ if $\frac{\alpha_B}{\alpha_N} = 1 - \sigma_N(L)$. Hence, $\sigma^0_N(L)^* = 1 - \frac{\alpha_B}{\alpha_N}$.

(ii) If $\alpha_B \geq \alpha_N$, then $\frac{\alpha_B}{\alpha_N} > 1$, and consequently, $\Delta^0_{N,L} = \Delta^0_{N,R} > 0$ for any $\sigma_N(L)$. Hence, the optimal strategy of the normal type is $(\sigma^0_N(L)^*, \sigma^0_N(R)^*) = (0,0)$. Now, by Lemma 3, we know that in this case, the wise type could find it optimal to use a strategy that does not always follow the agent's signal. For this case, Proposition 8 guarantees that the normal type has a unique equilibrium strategy, which is to always take action R.

To complete the proof, finally note that $\alpha_B \geq \alpha_N \iff \alpha_B \geq \hat{\alpha}_B = \frac{1-\alpha_W}{2}$. ♦

Lemma 15. *Suppose $\mu = 1$. Let $\bar{\alpha}^a_B = \frac{(2\gamma-1)(\gamma(1-\alpha_W)+\alpha_W)}{1-(1-\gamma)2\gamma}$ and $\bar{\alpha}^b_B = \gamma - (1-\gamma)\alpha_W$, with $\bar{\alpha}^a_B < \bar{\alpha}^b_B$.*

1. *If $\alpha_B \leq \bar{\alpha}^a_B$, there is a unique equilibrium. In the equilibrium, $\sigma^1_N(L)^* = 1$.*
2. *If $\alpha_B \in (\bar{\alpha}^a_B, \bar{\alpha}^b_B)$, there is a unique equilibrium. In the equilibrium, $\sigma^1_N(L)^* = \frac{\gamma(\alpha_W+\alpha_N)+(\gamma-1)(\alpha_B+\alpha_W)}{2\gamma\alpha_N(1-\gamma)} \in (0,1)$.*
3. *If $\alpha_B \geq \bar{\alpha}^b_B$, then $\sigma^1_N(L)^* = 0$ in the unique equilibrium strategy.*

Proof. First, note that

$$\Delta^1_{N,L} = (1-\gamma)\lambda_W(R,R) - \gamma\lambda_W(L,L) = \frac{(1-\gamma)\alpha_W}{\alpha_W+(\gamma+(1-\gamma)(1-\sigma_N(L)))\alpha_N} - \frac{\gamma\alpha_W}{\alpha_B+\alpha_W+\gamma\sigma_N(L)\alpha_N}, \tag{14}$$

with $\frac{\partial\Delta^1_{N,L}}{\partial\sigma_N(L)} > 0$. Consequently, there is either one root for $\sigma_N(L)$ or none. In other words, the equilibrium strategy is unique.

Now, note the following:

$\Delta^1_{N,L}\Big|_{\sigma_N(L)=1} \leq 0 \iff \alpha_B \leq \frac{(2\gamma-1)(\gamma+\alpha_W-\gamma\alpha_W)}{\gamma^2+(\gamma-1)^2}$. Let $\bar{\alpha}^a_B = \frac{(2\gamma-1)(\gamma+\alpha_W-\gamma\alpha_W)}{\gamma^2+(\gamma-1)^2}$. Consequently, $\sigma^1_N(L)^* = 1$ if $\alpha_B \leq \bar{\alpha}^a_B$.

$\Delta^1_{N,L}\Big|_{\sigma_N(L)=0} \geq 0 \iff \alpha_B \geq \gamma - \alpha_W(1-\gamma)$. Let $\bar{\alpha}^b_B = \gamma - \alpha_W(1-\gamma)$. Consequently, $\sigma^1_N(L)^* = 0$ if $\alpha_B > \bar{\alpha}^b_B$.

When $\bar{\alpha}^a_B < \alpha_B < \bar{\alpha}^b_B$, note that $\Delta^1_{N,L}\Big|_{\sigma_N(L)=0} < 0$ and $\Delta^1_{N,L}\Big|_{\sigma_N(L)=1} > 0$, which implies that there is one root. We obtain the following:

$\Delta^1_{N,L} = 0 \iff \sigma^1_N(L)^* = \frac{\gamma(\alpha_W+\alpha_N)+(\gamma-1)(\alpha_B+\alpha_W)}{2\gamma\alpha_N(1-\gamma)}$.

Finally, note that in the case $\alpha_B > \bar{\alpha}^b_B$, $\sigma^1_N(L)^* = 0$, which implies that the normal type uses a non-informative strategy. For this case, Lemma 3 says that the wise type could find it optimal to use a strategy that does not always follow the agent's signal. Finally, by Proposition 8, we know that in this case, the normal type has a unique equilibrium strategy, which is to always take action R. ♦

Lemma 16. *If $\sigma^\mu_N(L)^* \in (0,1)$, then $\sigma^\mu_N(L)^*$ is increasing in μ.*

Proof. First, note that $\Delta^\mu_{N,L} = (1-\mu)\Delta^0_{N,L} + \mu\Delta^1_{N,L}$, where from (13) and (14), we have $\frac{\partial\Delta^0_{N,L}}{\partial\sigma_N(L)} > 0$ and $\frac{\partial\Delta^1_{N,L}}{\partial\sigma_N(L)} > 0$. Thus, $\frac{\partial\Delta^\mu_{N,L}}{\partial\sigma_N(L)} > 0$, which means that the equilibrium strategy of the normal type is unique.

Now, let us denote by $\bar{\sigma}^0_N(L)^*$ and $\bar{\sigma}^1_N(L)^*$ the interior solutions to equations $\Delta^0_{N,L} = 0$ and $\Delta^1_{N,L} = 0$, respectively. From Lemmas 14 and 15 above, we know that these equilibrium strategies are $\bar{\sigma}^0_N(L)^* = 1 - \frac{\alpha_B}{\alpha_N}$ and $\bar{\sigma}^1_N(L)^* = \frac{\gamma(\alpha_W+\alpha_N)+(\gamma-1)(\alpha_B+\alpha_W)}{2\gamma\alpha_N(1-\gamma)}$.

Next, note that $\bar{\sigma}^0_N(L)^* < \bar{\sigma}^1_N(L)^* \iff 1 - \frac{\alpha_B}{\alpha_N} - \frac{\gamma(\alpha_W+\alpha_N)+(\gamma-1)(\alpha_B+\alpha_W)}{2\gamma\alpha_N(1-\gamma)} < 0 \iff \frac{(2\gamma-1)(\alpha_W(1-\gamma)\alpha_B+\gamma\alpha_N)}{2\gamma\alpha_N(1-\gamma)} > 0$, which is always the case.

Finally, note that $\Delta^0_{N,L} < 0$ if $\sigma_N(L) < \bar{\sigma}^0_N(L)^*$, and $\Delta^0_{N,L} > 0$ if $\sigma_N(L) > \bar{\sigma}^0_N(L)^*$. Similarly, $\Delta^1_{N,L} < 0$ if $\sigma_N(L) < \bar{\sigma}^1_N(L)^*$, and $\Delta^1_{N,L} > 0$ if $\sigma_N(L) > \bar{\sigma}^1_N(L)^*$. Then, if $0 < \sigma^\mu_N(L)^* < 1$, necessarily, $\sigma^\mu_N(L)^* \in \left[\bar{\sigma}^0_N(L)^*, \bar{\sigma}^1_N(L)^*\right]$. Finally, since $\Delta^\mu_{N,L} = (1-\mu)\Delta^0_{N,L} + \mu\Delta^1_{N,L}$, with $\Delta^0_{N,L} > 0$ and $\Delta^1_{N,L} < 0$ in the interval $\left[\bar{\sigma}^0_N(L)^*, \bar{\sigma}^1_N(L)^*\right]$, then $\sigma^\mu_N(L)^*$ has to be increasing in μ. ♦

This completes the proof of Proposition 9. □

References

Andina-Díaz, Ascensión, García-Martínez, José A., 2020. Reputation and news suppression in the media industry. Games Econom. Behav. 123, 240–271.

Ashworth, Scott, de Mesquita, Ethan Bueno, 2014. Is voter competence good for voters?: Information, rationality, and democratic performance. Am. Polit. Sci. Rev. 108 (3), 565–587.

Ashworth, Scott, Shotts, Kenneth W., 2010. Does informative media commentary reduce politicians' incentives to pander? J. Public Econom. 94 (11), 838–847.

Austen-Smith, David, Fryer, Roland G., 2005. An economic analysis of "acting white". Q. J. Econ. 120 (2), 551–583.

Avery, Christopher N., Chevalier, Judith A., 1999. Herding over the career. Econom. Lett. 63, 327–333.

Bagwell, Kyle, 2007. Signalling and entry deterrence: A multi-dimensional analysis. Rand J. Econ. 38 (3), 670–697.

Bar-Isaac, Heski, Deb, Joyee, 2014. (Good and bad) reputation of a servant of two masters. Amer. Econ. J.: Microeconom. 6 (4), 293–325.

Bénabou, Roland, Tirole, Jean, 2006. Incentives and prosocial behavior. Amer. Econ. Rev. 96 (5), 1652–1678.

Besley, Timothy, 2006. Principled Agents? The Political Economy of Good Goverment. Oxford University Press, Oxford.

Blumenthal, Benjamin, 2022. Political agency and implementation subsidies with imperfect monitoring. J. Law Econom. Organ. http://dx.doi.org/10.1093/jleo/ewac011.

Bourjade, Sylvain, Jullien, Bruno, 2011. The roles of reputation and transparency on the behavior of biased experts. Rand J. Econ. 42 (3), 575–594.
Canes-Wrone, Brandice, Herron, Michael C., Shotts, Kenneth W., 2001. Leardhip and pandering: A theory of executive policymaking. Amer. J. Polit. Sci. 45 (3), 532–550.
Devdariani, Saba, Hirsch, Alexander V., 2022. Voter Attention and Electoral Accountability. Working Paper.
Dewatripont, Mathias, Jewitt, Ian, Tirole, Jean, 1999. The economics of career concerns, part I: Comparing information structures. Rev. Econom. Stud. 66 (1), 183–198.
Ely, Jeffrey C., Välimäki, Juuso, 2003. Bad reputation. Q. J. Econ. 118 (3), 785–814.
Esteban, Joan, Ray, Debraj, 2006. Inequality, lobbying, and resource allocation. Amer. Econ. Rev. 96 (1), 257–279.
Feller, Miró, Schäfer, Ulrich, 2020. Deceiving Two Masters: The Effects of Capital and Labor Market Incentives on Reporting Bias. Working Paper.
Foerster, Manuel, Voss, Achim, 2022. Believe me, I am ignorant, but not biased. Eur. Econ. Rev. 149, 104262.
Fox, Justin, Shotts, Kenneth W., 2009. Delegates or trustees? A theory of political accountability. J. Polit. 71 (4), 1125–1137.
Fox, Justin, Van Weelden, Richard, 2012. Costly transparency. J. Public Econom. 96 (1), 142–150.
Frankel, Alex, Kartik, Navin, 2019. Muddled information. J. Polit. Econ. 127 (4), 1739–1776.
Gentzkow, Matthew, Shapiro, Jesse M., 2006. Media bias and reputation. J. Polit. Econ. 114 (2), 280–316.
Gersbach, Hans, Hahn, Volker, 2008. Should the individual voting records of central bankers be published? Soc. Choice Welf. 30 (4), 655–683.
Holmström, Bengt, 1999. Managerial incentive problems: A dynamic perspective. Rev. Econom. Stud. 66 (1), 169–182.
Hörner, Johannes, 2002. Reputation and competition. Amer. Econ. Rev. 92 (3), 644–663.
Leaver, Clare, 2009. Bureaucratic minimal squawk behavior: Theory and evidence from regulatory agencies. Amer. Econ. Rev. 99 (3), 572–607.
Levy, Gilat, 2005. Careerist judges and the appeals process. Rand J. Econ. 36 (2), 275–297.
Levy, Gilat, 2007. Decision making in committees: Transparency, reputation, and voting rules. Amer. Econ. Rev. 97 (1), 150–168.
Li, Ming, Madarász, Kristóf, 2008. When mandatory disclosure hurts: Expert advice and conflicting interests. J. Econom. Theory 139 (1), 47–74.
Liu, Yaozhou Franklin, Sanyal, Amal, 2012. When second opinions hurt: A model of expert advice under career concerns. J. Econ. Behav. Organ. 84 (1), 1–16.
Maskin, Eric, Tirole, Jean, 2004. The politician and the judge: Accountability in government. Amer. Econ. Rev. 9 (4), 1034–1054.
Morris, Stephen, 2001. Political correctness. J. Polit. Econ. 109 (2), 231–265.
Ottaviani, Marco, Sørensen, Peter N., 2001. Information aggregation in debate: Who should speak first? J. Public Econom. 81 (3), 393–421.
Ottaviani, Marco, Sørensen, Peter N., 2006. The strategy of professional forecasting. J. Financ. Econ. 81 (2), 441–466.
Prat, Andrea, 2005. The wrong kind of transparency. Amer. Econ. Rev. 95 (3), 862–877.
Sibert, Anne, 2003. Monetary policy committees: Individual and collective reputations. Rev. Econom. Stud. 70 (3), 649–665.